International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 18 Issue 1 – OCTOBER 2015.

MC- CSO: PERFORMANCE ANALYSIS AND DETECTING THE INTRUSION DETECTION SYSTEM USING SWARM TECHNIQUE

B.Subbulakshmi[#1], E.Ramaraj[*2]

[#1]*Research Scholar, Madurai Kamaraj University*
[*2]*Professor, Department of Computer Science and Engineering, Alagappa University, Karaikudi*

***Abstract*--Intrusion Detection System plays a crucial role in security oriented systems that tracks and records the real time data which leads to dimensionality reduction, multiclass oriented functionality dependency and symmetric issue. In this paper, we proposed Multiple Criteria – Cat Swarm Optimization (MC- CSO) which lessened the dimensionality reduction, multiclass oriented functionality dependency and symmetric issue. The widely known 10% KDD cup 99 dataset is studied in this paper. The inability of detecting new patterns of network traffic is the demerit found in existing technologies. The multiple criteria system belongs to the class of Linear Programming Model (LPM). The velocity and position of cat is updated using multiple criteria system. Experimental results were shown that our novel technique works efficiently than the existing systems in terms of higher accuracy rate and lessened false positive rate.**

Keywords: Intrusion Detection System, Symmetric issue, Linear Programming Model, Cat Swarm Optimization.

I. INTRODUCTION

Data mining defined as obtaining or mining a valid knowledge from enormous amount of information [1]. Data mining is also known as ‘knowledge mining from data ‘or ‘knowledge mining’. These types of data collection and its storage application generate a greatest way for enterprises at lower cost. Employing these types of archived facts to get valid and executable information is the main goal of the data mining. The data mining tasks listed as [2]:

i) ***Exploratory Data Analysis:*** This data analysis possesses a large amount of data in its repositories. It works on two purposes:

a) Without the knowledge for what the customer is looking for.

b) It explores the data that are interactive and visual to the customer.

ii) ***Descriptive Modeling****:* It explains all about the data. It has a probability distribution of the data. It also conveys a relationship between the variables [2].

iii) ***Predictive Modeling****:* This defines a new model based on the past information model [3].

iv) ***Discovering patterns and rules****:* This role is to predict a new pattern as well as hidden pattern in the cluster [4]. It has different types of patterns with different size. The aim of the task is ‘how far the best patterns are detected’. Many clustering algorithms discovered to get new patterns and rules.

v) ***Retrieval by Content****:* The aim of this task is to discover the sets of data. It obtains a similar pattern of interest in the data set [5].

Network security turns out to be more essential with the gigantic development of computer system utilization both interior and exterior. To shield against different assault, a loads of computer security systems have been seriously contemplated in the most recent decade. Among them Network Interruption Identification (NII) has been thought to be a standout amongst the most favorable techniques for protecting and anomaly detection. The detection of misuse behavior is known Signature based Intrusion Detection System (IDS). It has ability to process huge measure of information and decrease the information and by separating particular information, which enhances execution streamlining of locating the rules. The algorithms studied were the Bayesian technique [1, 29] and decision tree [2, 3]. In IDS, the detection rate is of high and low false alarm rate is a challenging task. Lately numerous science propelled approach, for example, Genetic Algorithm (GA) [6] [7], Genetic Programming (GP), Ant Colony (AC) [8], Immune Algorithm, Artificial Bee Colony and Swarm Intelligence (SI) [9] assumes to be a crucial part in IDS to enhance their productivity and execution.

This paper is organized as Section I depicts the basic needs in Data mining and Network Security system. Section II depicts the related study carried out by various researchers. Section III portrays the proposed technique; Multiple Criteria based Cat Swarm Optimization (MC-

CSO). Experimental study is carried out in Section IV. Atlast concluded in Section V.

II. RELATED WORK

The raw network information has enormous system traffic dataset which prompts immaterial and enormous dimensionality issue. The KDD'99 [12, 15] has been the most uncontrollably utilized information set for the assessment of Intrusion Detection System (IDS) is studied by Stolfo et al. Agarwal and Joshi [13] proposed a two-stage general-to-particular system for researching tenet based model (PNrule) to learn classifier models on an information set that has broadly distinctive class disseminations in the training data. The proposed PN tenet assessed on KDD dataset reports high detection rate. Yeung and Chow [14] proposed a curiosity identification methodology utilizing no-parametric density estimation in view of Parzen-window estimators with Gaussian threads to construct an interruption recognition framework utilizing typical information. This anomaly identification methodology was utilized to recognize assault classes in the KDD dataset. In 2006, Xin Xu et al. [17] exhibited a system for versatile interruption location in view of machine learning. Multi-class Support Vector Machines (SVMs) is connected to classifier development in IDSs and the execution of SVMs is assessed on the KDD99 dataset. In [18], Portnoy et al. divided the KDD information set into ten subsets, each containing roughly 490,000 occasions or 10% of the information. Then again, they noticed that the appropriation of the assaults in the KDD information set is extremely uneven which made cross-approval system for smurf and Neptune based assaults. The natural roused methodologies have been widely incited in Intrusion Detection System. For instance, another collaborative system for pre-handling the test sort of assaults is proposed by G. Sunil Kumar, [19] in light of cross breed classifiers on double particle swarm improvement and random forest estimation for the grouping of test assaults in a system. Wei-Chang yeh et.al [20] proposed new system by consolidating SSO with weighted trade neighborhood look strategy for interruption identification.

III. MULTIPLE CRITERIA BASED CAT SWARM OPTIMIZATION (MC- CSO)

Multiple Criteria based Cat Swarm Optimization (MC- CSO) is the technique used for detecting the network traffic pattern. The multiple criteria system belongs to the class of Linear Programming Model (LPM). This scheme has been proposed by Shi in 1998. This is one of the best optimization classification systems. Let us assume data of class as C= {C_1, C_2, …C_n}. Consider a linear separable hyperplane.

$$w.c = d$$
(3.1)

Where w is the weight of the subsets of the Class C and b is the user-defined threshold. Hence, this classifier attempt to locate the best hyper-plane which can precisely isolate the class from the samples and also predict the new data belongs to the specified class or not. Assume two class label, N as Normal and A as Attack. The linear hyperplane is subdivided into two groups based on two class label.

$$(w.c_i) \leq d; \forall c_i \in N$$
(3.2)

$$(w.c_i) \geq d; \forall c_i \in A$$
(3.3)

The classification problem is handled by two objective functions. The samples may lead to an intersection of any data point in hyperplane. The first objective is to reduce the aggregate of intersecting points in linear hyperplane. It is known as LAD (Lower bound of aggregation of Distance). The second objective is to maximize the LAD which is known as MLD (Maximizing the Lower bound of Distance aggregation). In the MCLP system, the two parameters are assumed. Let α_i denotes the intersecting points with reference to training sample C_i which should be minimized. Let β_i denotes maximizing the distance from training sample Ci to the hyperplane. The conditions are formulated as:

$$(w.c_i) \leq d + \alpha_i - \beta_i, \forall c_i, y = Normal$$
(3.4)

$$(w.c_i) \geq d - \alpha_i + \beta_i, \forall c_i, y = Attack$$
(3.5)

Based on the objectives, the MCLP is given as $\min \sum_{i=1}^{n} \alpha_i$ and $\max \sum_{i=1}^{n} \beta_i$ where $\alpha, \beta \geq 0$

Thus, the Multiple Criteria Linear Function is formed. This MCLP is merged with the Cat Swarm Optimization technique. Swarm Intelligence is one of the best yielding optimal solutions. Some complicated real time problems can be easily solved by the Swarm Intelligence. There are many types of swarm systems namely, behavior of ants, bird flocking, fish schooling and cats. This paper intends to study about the Cat Swarm Optimization technique. The advantages of using CSO were the Versatility, Speedier Execution and Faster in predicting approximate solutions. In CSO, the network users are denoted as the Cats. After the parameter initialization, each cat updates its velocity and position in every iteration. Based on the knowledge extracted from the previous iterations, the cats are processed to the seeking mode and discovering the best solutions in tracing mode. At the end, the best solution is inferred from Velocity and Position equation as follows:

$$V_c^i = w.V_c^i + r_1 * c_1 * [x_{pbest}^c - x_c^i] \quad (3.6)$$

Where c= 1, 2….N, N is the number of cats population. w is the weight vector which is used for global and local search. i denotes the iteration level of a cat. Pbest is the best position of the cat c at the iteration i. r_1 denotes the random variable range [0, 1] and C_1 denotes the constant which takes the value 2.05. The position is estimated as:

$$P_c^i = x_c^i + V_c^i \quad (3.7)$$

IV. EXPERIMENTAL RESULTS

Stolfo et al machinated the KDD CUP 99 dataset from the IDS evaluation program DARPA'98. DARPA 98 comprised of raw data of 7weeks of network traffic of about 4 gigabytes. Then it is preprocessed into 5 million connection records of 100 bytes. The test data consists of 2 million connection records monitored of two weeks of network traffic. Similarly, the training dataset consists of 41 features of 4,900,000 connection vectors under the labeled classes either NORMAL or an ATTACK with the specific type of attack. This type of dataset is widely used in the field of Data stream classification to find the anomaly or refreshed classes. The categorized types of attacks were the following:

1) ***Denial of Service attack (DoS):*** Some intruders intend to disaster the server machine by giving sequence of requests, unauthorized users access or creating less resistance power for fault tolerant systems.
2) ***User to Root attack (U2R):*** It taps the task of normal user accounts such as sniffing passwords, dictionary attack or any social engineering practices to create degree of vulnerability issues.
3) ***Remote to Local Attack (R2L):*** The intruders send packets to the any node in the system to create susceptibility over the network to access the information of the users.
4) ***Probing attack:*** The intruders destined to collect the entropy of computers and its networks to discover the security controls.

The KDD'99 data can be organized into three features:

1. ***Basic features:*** The features of TCP/IP connection are collected. The inherent features of delay in detection are encapsulated as 'Basic features'.
2. ***Traffic features:*** The window interval features are collected and categorized into two forms: ***a) "Same host" features:*** It records the current connection of past 2 seconds that include the same destination host and its protocol behavior, service etc are estimated for statistical performance.
   b) "Same service" features: It records the connection information of past 2 seconds that holds the same service. The traffic features is a time- series based. Some probing attacks use the host and service for more than two seconds which results to the non-intrusion patterns. This issue can be solved by the 'connection based traffic features' which utilizes the records of 100 connections of the connection window.
3. ***Content features:***

In contrast to the DoS and probing attacks, the R2L and U2R do not possess any intrusion patterns as it embeds the data segments of the packets. To discover these types of attacks, some features has to be monitored e.g. number of failed login attempts, source host rate, destination host rate etc. And these are known as 'content features'.

The training dataset consisted of 65,536 records among which 6, 368 (9.17%) were normal, 57,266 (83.38%) DOS, 809 (1.23%) Probe, 1,090 (1.66%) R2L and 3 (0.004%) U2R connections. In each connection, there are 41 attributes describing different features of the connection and a label assigned to each either as an attack type or as normal. Table 1 shows the class labels and the number of samples that appears in "10% KDD" training dataset.

Table 1: Class labels and the number of samples that appears in 10% KDD dataset

Attack	Original number of samples	Class
Apache2	398	DOS
Buffer_overflow	3	U2R
Guess-Password	2	R2L
ipsweep	80	PROBE

Multihop	4	R2L
Named	10	DOS
Normal	6368	NORMAL
Phf	1	R2L
pod	22	DOS
portsweep	106	PROBE
saint	623	PROBE
Sendmail	6	R2L
Smurf	56835	DOS
snmpgetattack	1069	R2L
udpstorm	1	DOS
xlock	6	R2L
xsnoop	2	R2L
Total	65536	

Table 2: Protocol Type in our dataset

Protocol Type	Attack Type
UDP	2705
TCP	5867
ICMP	56960

As far as we know for better intrusion detection, the detection rate must be higher and false alarm rate must be lower. To analysis the system performance, the parameters such as accuracy, novel class detection rate and false alarm rate are estimated.

Table 3: No. of data in variants attacks

Attacks Type	**No. of records**
DOS	57266
U2R	3
R2L	1090
PROBE	809
NORMAL	6368

In this proposed MC- CSO model, a simple classification is made as the training data sets that contain only normal and a DOS attack is selected randomly to detect the intrusion creating users. The proposed step is as follows:

1. Consider a data of class with n-dimensional attribute. The training sample is assumed as C= $\{c_1, c_2 \ldots c_n\}$. Assume the y vector that include the range [-1, +1] where -1 represents the normal system and -1 represents the DOS attacks.
2. The CSO parameters are defined as w, r_1, c_1, n, α and β . The values are initialized as r_1= [0, 1]; c_1= 2.05; n=30; w_{max}= 0.85; w_{min}=0.30; α be the lower bound that denotes negative integer between 0.00001 to -0.1. β be the upper bound that denotes positive integer between 10 to 10000.
3. The accuracy of the model that dictates the prediction of corrected records over the total records. It is given as $Acc = \frac{(TP + TN)}{TP + FP + FN + TN}$ and False Alarm Rate $FAR = \frac{FP}{FP + TN}$
4. Based on the updated velocity and position of cat in eqn. 3.6 &3.7.
5. Stop the process until the highest accuracy and optimal value from α and β is achieved.

The confusion matrix is given as:

Confusion matrix		Predicted class label	
		Normal	Intrusions
Observed class label	Normal	True negative	Error rate of misclassified instances
	Intrusions	False negative	Correctly detected class.

Table 4: Comparison results with the existing classifier.

Classifier	**Accuracy in %**
Simplified Particle Swarm Optimization	98.7%
Support Vector Machines	65.1%
Naïve Bayes	88.6%
Random Forest	92.7%
MC- CSO	98.9%

V. CONCLUSION

This paper depicted the performance validation of our proposed technique named; **Multiple Criteria based Cat Swarm Optimization (MC- CSO)** using 10% KDD cup dataset. The MC-CSO systems can easily lessened the dimensionality reduction, multiclass oriented functionality dependency and symmetric issue. The dataset contain only the most relevant feature subset which portrayed with two class label as Normal and Attack to form network traffic. In this MC- CSO technique, the normal records were eliminated, thereby to reduce the work of Intrusion Detection System (IDS). When related to other data mining systems, the swarm oriented data mining systems can efficiently and effectively produces an optimal solution with higher accuracy rate. With the approximated optimal solution, the rate of false positive is reduced. The performance analysis proves that the MC- CSO technique can easily generate the optimal solution and more efficient in solution.

REFERENCES

[1] John, G.H., Langley, P: “Estimating Continuous Distributions in Bayesian Classifiers”. In Proc. of the 11th Conf. on Uncertainty in Artificial Intelligence 1995.

[2] Quinlan, J.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo 1993.

[3] Kohavi, R.: “Scaling up the accuracy of naïve-bayes classifier: A decision-tree hybrid”. In: Proc. Of the 2nd International Conference on Knowledge Discovery and Data Mining, pp. 202–207. AAAI Press, Menlo Park 1996.

[4] Witten, I.H., Frank, E.: Data Mining: “Practical Machine Learning Tools and Techniques”, 2nd edn. Morgan Kaufmann, San Francisco 2005.

[5] Werbos, P.: Beyond Regression: “New Tools for Prediction and Analysis in the Behavioral Sciences”. PhD Thesis, Harvard University (1974)

[6] Al-Tabtabai H, Alex PA. “Using genetic algorithms to solve optimization problems in construction”. Eng Constr Archit Manage 1999; 6(2):121–32.

[7] Dharmendra G. Bhatti, P. V. Virparia, Bankim Patel, “Conceptual Framework for Soft Computing based Intrusion Detection to Reduce False Positive Rate’, International Journal of Computer Applications (0975 – 8887), Volume 44– No13, April 2012.

[8] Dorigo M, Di Caro G. 1999. “The ant colony optimization meta-heuristic, new ideas in Optimization”, ACM Transaction, 11-32.

[9] Yao Liu, Yuk Ying Chung, and Wei-Chang Yeh: “Simplified Swarm Optimization with Sorted Local Search for golf data classification”. IEEE Congress on Evolutionary Computation 2012: 1-8

[10] Deris tiawan, Abdul Hanan Abdullah, Mohd. Yazid dris, “Characterizing Network Intrusion Prevention System”, International Journal of Computer Applications (0975 – 8887), Volume 14– No.1, January 2011.

[11] Kennedy J, Eberhart R. “Particle swarm optimization”. Proceedings of the IEEE international conference on neural networks (Perth, Australia), 1942–1948. Piscataway, NJ: IEEE Service Center; 1995.

[12] KDDCUP 99 dataset, available at: http://kdd.ics.uci.edu/dataset/kddcup99/kddcup99.html.

[13] Agarwal, R., Joshi, and M.V.: PNrule: “A New Framework for Learning Classifier Models in Data Mining”. Tech. Report, Dept. of Computer Science, University of Minnesota 2000.

[14] Yeung, D.Y., Chow, C.: Prazen-“window Network Intrusion Detectors”. In: 16th International Conference on Pattern Recognition, Quebec, Canada, pp. 11–15 August 2002.

[15] S. J. Stolfo, W. Fan, W. Lee, A. Prodromidis, and P. K. Chan, “Cost based modeling for fraud and intrusion detection: Results from the jam project,” discex, vol. 02, p. 1130, 2000.

[16] MIT Lincoln Labs, 1998 DARPA Intrusion Detection Evaluation. Available on: http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/ index.html, February 2008.

[17] Xu, X.: “Adaptive Intrusion Detection Based on Machine Learning: Feature Extraction, Classifier Construction and Sequential Pattern Prediction”. International Journal of Web Services Practices 2(1-2), 49–58 2006.

[18] L. Portnoy, E. Eskin, and S. Stolfo, “Intrusion detection with unlabeled data using clustering,” Proceedings of ACM CSS Workshop on Data Mining Applied to Security, Philadelphia, PA, November, 2001.

[19] G. Sunil Kumar, C.V.K Sirisha, Kanaka Durga.R, A.Devi, “Robust Preprocessing and Random Forests Technique for Network Probe Anomaly Detection”, International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Volume-1, Issue-6, and January 2012.

[20] Changseok bae, Wei-Chang yeh, Noorhaniza Wahid, yuk ying chung and yao liu, “A new simplified swarm optimization (SSO) using exchange local search scheme”. ICIC International @ 2012. Volume 8, number 6, June 2012.

[21] Dharmendra G. Bhatti, P. V. Virparia, Bankim Patel, “Conceptual Framework for Soft Computing based Intrusion Detection to Reduce False Positive Rate”, International Journal of Computer Applications (0975 – 8887), Volume 44– No13, April 2012.

[22] Ian H. Witten, Eibe Frank “Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)”, Morgan Kaufmann June 2005, 525 pages Paper ISBN 0-12-088407-0.

[23] C. Romero, S. Ventura and E. García, “Data mining in course management systems: Moodle case study and tutorial”, Computers & Education, Volume 51, Issue 1, pp. 368-384, 2008, Elsevier Science.

[24] S. H. Zahiri and S. A. Seyedin, “Swarm intelligence based classifiers”, Journal of the Franklin Institute, vol.344, no.5, pp.362- 376, 2007.

[25] Yao Liu, Yuk Ying Chung, and Wei-Chang Yeh: “Simplified Swarm Optimization with Sorted Local Search for golf data classification”. IEEE Congress on Evolutionary Computation 2012: 1-8.

[26] K. Shafi, H.A. Abbass, “Biologically inspired complex adaptive systems approaches to network intrusion detection”, Information Security Technical Report 12 (4) ,2007, 209–217.

[27] Bosh, A., Zisserman, A., Munoz, and X.: "Image classification using Random Forests and ferns". In: IEEE ICCV 2007.

[28] Ned Horning, "Introduction to Decision Trees and Random Forests", American Museum of Natural History's.

[29] G. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, pp. 338–345, 1995.

[30] R. Kohavi, "Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid," in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, vol. 7, 1996.

[31] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[32] C. Chang and C. Lin, "LIBSVM: a library for support vector machines," 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[33] "Waikato environment for knowledge analysis (weka) version 3.5.7." Available on: http://www.cs.waikato.ac.nz/ml/weka/, June, 2008.