Impression Intrusion Detection System Using Fuzzyset Clustering Algorithm

Bh. Dasaradha Ram[1], Dr. B.V. Subba Rao[2]
[1] Dept of CSE
[2]HOD, Dept of IT
[1]Rayalaseema University, Kurnool, (AP) – India
[2]pvpsiddhartha Engineering College (AP) – India

Abstract- *Intrusion detection is one of the major fields of research and researchers are trying to find new algorithms for detecting intrusions. Clustering techniques of data mining is an interested area of research for detecting possible intrusions and attacks. The present dectection method different overview of existing Intrusion Detection Systems (IDS) along with their main principles. We propose new intrusion detection system based on a parallel particle swarm optimization clustering algorithm using the MapReduce approaches. In our proposed framework for a Parallel Fuzzy Genetic Algorithm (PFGA) is developed classification and prediction over decentralized data sources. The model parameters are evolved using two nested genetic algorithms (GAs). The outer GA evolves the fuzzy sets whereas the inner GA evolves the fuzzy rules. During optimization, best rules are only distributed among agents to construct the overall optimized model. We implement our experiments K-means clustering algorithm and measured the performance based on detection rates and false positive rate with different cluster values. The KDD dataset which is freely available online is used for our experimentation and results are compared.*

Keywords- Network, Attacks, k-means Clustering, Securitym, Fuzzy Classification; Rule-Base; Fuzzy Logic System (FLS); Genetic Algorithm; Distributed Data Mining (DDM)

I. INTRODUCTION

Intrusion Detection System (IDS) is a device typically another separate computer that monitors activity to identify malicious or suspicious events [1]. Traditionally network users usually use firewall as the first line of defense for security. But with attacking tools and means becoming much more complicated, simple firewall is difficult to resist various attacks [2]. On the other hand, anomaly-based IDSs are able to detect new attacks that have not been seen before. However, this model produces a large number of false positives. The reason for this is the inability of current anomaly-based techniques to cope adequately with the fact that in the real world, normal, legitimate computer networks, and system usage changes over time [3]. This implies that any profile of normal behavior needs to be dynamic [4]. The pattern match against packet in network for worm signature detection [5]. We have used KDD dataset. Additionally the features were reduced to some level to have better accuracy using principal component analysis and ranking algorithms [6]. New detection has a key advantage is their high rate of speed of detecting known attacks. Their main drawback is the inability to detect many attacks. Anomaly detection built profiles based on normal behavior [7]. Anomaly detection new attacks hybrid intrusion detection system combine the advantages of misuse and anomaly detection [8]. New classifiers fezzy set of IF-THEN rules for classification [9]. Rule-based classification model is shortcoming that they involve sharp cutoffs for continuous attributes. Fuzzy Logic System (FLS) is attractive features that make it an alternative to designing data mining systems performing rule-based classification effectively data sets [10]. The most important objective of this paper is to evaluate performance of k means algorithm for design of clustering based intrusion prevention system. Therefore compare the performance of k-means algorithm with different values of clusters. We investigated the performance of k-means clustering algorithm in terms of performance criteria as detection rate and false positive rate [11].

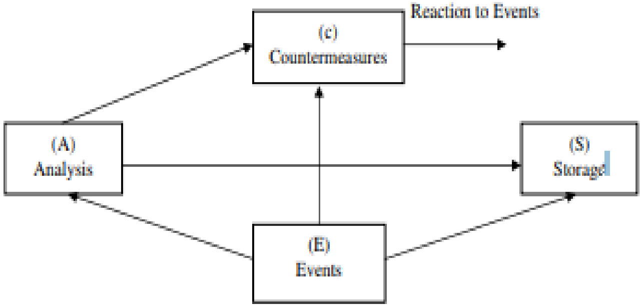

Figure 1. Common components of an Intrusion Detection Framework

II. RELATED WORK

Unsupervised anomaly detection techniques detect anomalies in an unlabelled data set assuming that the maximum numbers of instances in the dataset are normal training data. The techniques in this category make the implicit assumption that normal instances are more frequent than anomalies in the test data [12]. Two data mining techniques are random forest and k-means algorithms are used in misuse anomaly and hybrid detection [13]. However in order to evaluate candidate solutions in inner GA a chromosome from the outer GA must be utilized since it encodes the fuzzy sets definitions required in evaluating the rule base encoded to modify the security of classification or prediction [14]. In this case the fitness function is simply defined as the testing error. The artificial immune system is designed for the computational system and inspired is applied to solving many problems in the field of information security particularly intrusion detection systems [15]. They decomposed the normal data into smaller subsets using misuse detection model. The SVM models were built many decomposed data to have precise behavior from the normal data profile. They claim that the model outperforms conventional methods in terms of detection rate and low false positive rate [16]. Anomaly detection based intrusion detection systems work based on a profile of a normal network or system using statistical machine learning techniques. Anomaly detection based on machine learning techniques can be categorized as either supervised unsupervised depending on whether the class labels are known during the learning process. Several techniques is proposed to tackle the intrusion detection problem using unsupervised algorithms like clustering-based algorithms [17]. The proposed SVM classification is formed and lastly classification using radial SVM is performed to detect intrusion has happened or not [18].

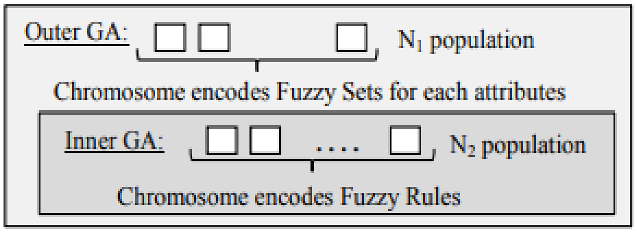
Outer GA:
N₁ population
Chromosome encodes Fuzzy Sets for each attributes
Inner GA:
. . . .
N₂ population
Chromosome encodes Fuzzy Rules

Fig. 2. The structure of nested GAs that evolves local model parameters of FGA agent

III. SYSTEMS MODEL

Map Reduce usually divides the input data set into independent splits which depend on the size of the data set and the number of computer nodes used. Map Reduce consists of two main functions: Map and Reduce functions [19]. The Map function processes the input data records as (key, value) data pairs to generate intermediate output as (key, values list) data pairs and then the Reduce function merges and aggregates all intermediate (values list) output coming from the Map function having the same intermediate key [20]. Particle swarm optimization (PSO) algorithm which is evolutionary computation technology based on swarm intelligence has good global search ability. Experiments on data sets KDD CUP 99 is effectiveness of the proposed method and also shows the method has higher detection rate and lower false detection rate is explained the process of intrusion detection which is the major part of network activity and security policies adapted over the network to secure it

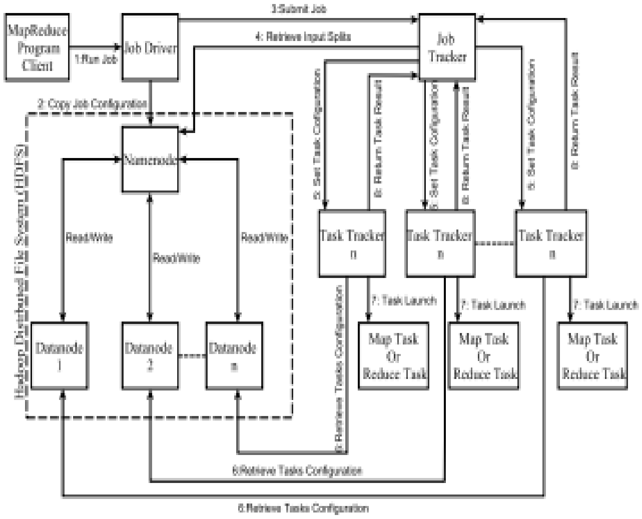

Fig. 3. Hadoop Architecture Diagram

IV. PROPOSED SYSTEM

Given that our proposed intrusion detection system is based on PSO clustering using Map Reduce methodology, we first briefly introduce PSO, introduce PSO clustering using Map Reduce [21], and then outline the details of our proposed intrusion detection system. The purpose of this process is to use only numerical data in our distance calculation because for the categorical data the distance calculations are difficult and depend on the data itself [22]. We discuss the K-means algorithm in this section. K-means is an iterative clustering algorithm in which items are processed among set of clusters until the desired set is reached. K-means algorithm is like a squared error algorithm [23].

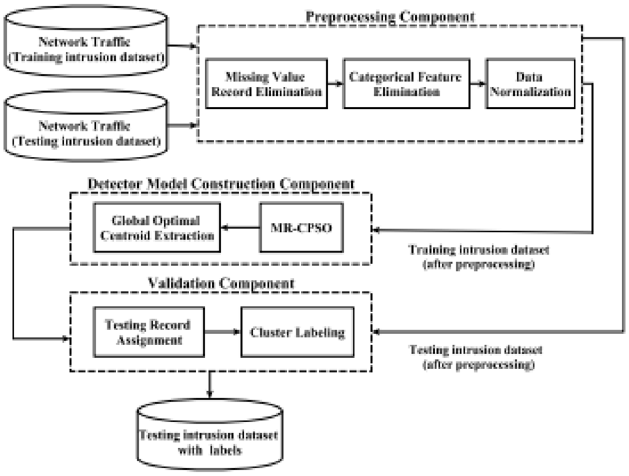
Preprocessing Component
Network Traffic (Training intrusion dataset)
Network Traffic (Testing intrusion dataset)
Missing Value Record Elimination
Categorical Feature Elimination
Data Normalization
Detector Model Construction Component
Global Optimal Centroid Extraction
MR-CPSO
Training intrusion dataset (after preprocessing)
Validation Component
Testing Record Assignment
Cluster Labeling
Testing intrusion dataset (after preprocessing)
Testing intrusion dataset with labels

Fig. 4. Proposed IDS-MRCPSO Architecture Diagram

A. PROPOSED ALGORITHM

The k-medoids algorithm and its modifications are used. The k-medoids algorithm is also a partitioning technique of clusters that clusters the data sets of n objects into k clusters with apriori. It could be more robust to noise and outliers as compared to K-means since it minimize a sum of pair wise dissimilarities using a squared Euclidean distance [24].

New Medoid Clustering Algorithm

Input: D dataset of n object

Output: Desired set of normal and abnormal clusters.

Begin

Step1: Standardize the dataset in order to make the feature value to appropriate range. This is done because features with greater value dominate the features with lesser value.
Step2: Select initial medoids and for that the formula of Euclidean distance for dissimilarity measure has been used. It is given as under
Step3: Associate each object to its closest medoid and calculate the optimal value as the sum of distances from all objects to their medoids.
Step4: Swap the current medoid in each cluster by the object which minimizes total distance to other objects in the cluster.
Step5: Again associate each object to the closest medoids and compute the new value as in step3. If the new value is same as previous one then stop the algorithm otherwise repeat step4.

End

The above algorithm will result in cluster formation and the next steps is to check for an empty cluster, if there is an empty cluster then remove the empty cluster by deleting them and hence this will eliminate degeneracy problem.

B. CLUSTERING ALGORITHMS

To create clusters from the input data, we have used k-means clustering algorithm is well-known clustering problem. The algorithm initially have empty set of clusters and updates it as proceeds. For each record it computes the Euclidean distance between it and each of the centroids of the clusters. Assume we have fixed metric M, and constant cluster Width W. Let di(C, d) is the distance with metric M, Cluster centroid C and instance d where centroid of cluster is the instance from feature vector [25].

Input: The number of clusters K and a dataset for intrusion detection

Output: A set of K-clusters

Algorithm:

1. Initialize Set of clusters S.
2. While cluster structure changes, repeat from 2.
3. Determine the cluster to which source data belongs Use Euclidean distance formula. Select d from training set. If S is empty, then create a cluster with centroid as d. else add d to cluster C with min (dist (C, d)) or dist(C ,d)<=dist(C1, d).
4. Calculate the means of the clusters. Change cluster centroids to means obtained using Step 3.

V. FUZZY LOGIC SYSTEMS (FLSS)

One highly successful theory in Computational Intelligence (CI) techniques is fuzzy set theory [26]. The design of FLS was one of the largest application areas derived from fuzzy set theory. FLS have demonstrated their superb ability as system identification tools and has enjoyed wide popularity in computer science and engineering as an advanced Artificial Intelligence (AI) tool and control technique [15] [16]. Recent work by data mining researchers has shown that the qualitative nature of FLS makes it a formal tool for constructing classifiers that deal with problems characterized by pervasive presence of uncertainty. For example, Fuzzy-based classifier has been applied successfully in data mining for Hepatitis [19], and data mining for intrusion detection [20]. Fuzzy-based classifier, generally, consists of a set of fuzzy linguistic rules as sentences rather than equations.

These fuzzy linguistic rules are easier understood than systems of mathematical equations. A FLS, generally, is known as knowledge-based system. The Knowledge Base (KB) not only has the rule-base but it also has the fuzzy sets and membership functions of the fuzzy partitions associated to the linguistic input and output variables. Therefore, this specifies a clear distinction between the fuzzy model structure and parameters as defined in classical knowledge discovery techniques [21].

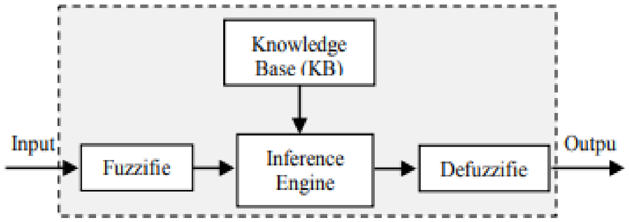

Fig. 5. The structure of a Fuzzy Logic System (FLS)

VI. EXPERIMENTAL RESULT

The KDD cup99 data set used as an input in order to conduct our experiment and later to check the performance of the algorithm The kddcup99 dataset is the most commonly used dataset for intrusion detection first given by Massachusetts Institute of Technology. Standardization of the dataset has been done so that it would be appropriate to be used by the proposed algorithm. The proposed algorithm is compared against the existing algorithm on the basis of parameters detection rates, accuracy and false alarm rate. he proposed algorithm is compared against the existing algorithm on the basis of parameters detection rates, accuracy and false alarm rate. In order to improve detection rate and accuracy a modified version of k-medoid algorithm is used and this eliminates the disadvantages of K-means algorithm. Also the degeneracy is being eliminated by using the concept of searching for the empty clusters and deleting them.

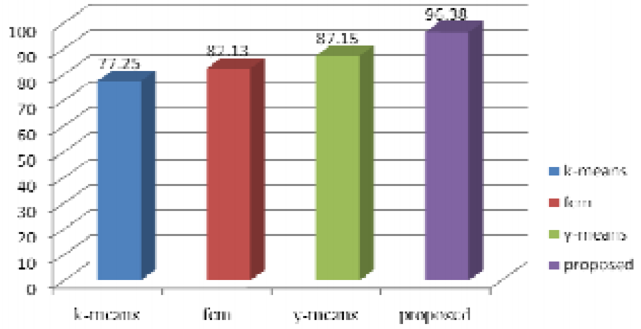

Figure 6. Comparisons of Detection Rate against various attacks

VII. CONCLUSION AND FUTURE WORK

The algorithm specified a new way of selection of initial medoid and proved to be better than K-means for anomaly intrusion detection. The algorithm conveys the idea of data mining technologies which is certainly a good field and popular area of research in intrusion detection. The intrusion detection system can be parallelized efficiently with the MapReduce methodology. Experiments were performed on a real intrusion data set in order to measure the system speedup. The developed Parallel Fuzzy-Genetic Algorithm (PFGA) framework provides flexible mechanism for processing distributed data and offers significant advantage over classical techniques which help to reach all network-related business. Our future work involves development of intrusion protection system to achieve low false positive rate and more accuracy using anomaly based detection approach. we have the plan to implement an online NIDS which can provide real-time feedback to the system, so that the unintentional delay from the offline detection method can be eradicated.

REFERENCES

[1] Ulf lindvist, Phillip Brentano and Doug Mansur, "IDS Motivation, architecture, and An Early Prototype", Computer Security Laboratory, US Davis: 160-171\

[2] McHugh, “Testing intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory,” ACM Transactions on Information and System Security, vol. 3, no. 4. pp. 262-294, 2000.

[3] M. Tavallaee et al. “A detailed analysis of the KDD CUP 99 data set,” in Proc. the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications, 2009.

[4] G. Kim, S. Lee, and S. Kim. “A novel hybrid intrusion detection method integrating anomaly detection with misuse detection,” Expert Systems with Applications, vol. 41, no. 4. pp. 1690-1700, 2014.

[5] B. Luo and J. Xia, “A novel intrusion detection system based on feature generation with visualization strategy,” Expert Systems with Applications, 2014.

[6] C. J. Fung and R. Boutaba, “Design and management of collaborative intrusion detection networks,” in Proc. 2013 IFIP/IEEE International Symposium on Integrated Network Management, 2013.

[7] K. Alsubhi, M. F. Zhani, and R. Boutaba, “Embedded Markov process based model for performance analysis of Intrusion Detection and Prevention Systems,” in Proc.

IEEE Global Communications Conference (GLOBECOM), 2012

[8] Reda M. Elbasiony, Elsayed A. Sallam, Tarek E. Eltobely, Mahmound M. Fahmy, “A hybrid network intrusion framework based on random forest and weighted k-means,” Ain Shams Engineering Journal, 2013.

[9] S. Forrest, S. A. Hofmeyr, and A. Somayaji, “Computer Immunology,” Commun. ACM, vol. 40, no. 10, pp. 88–96, Oct. 1997.

[10] A. Fariz, J. Abouchabaka, and N. Rafalia, “Using multi-agents systems in distributed data mining: a survey”, in Journal of Theoretical & Applied Information Technology, 73(3), 2015.

[11] S. V. S. G. Devi, “A survey on distributed data mining and its trends”, in IMPACT: International Journal of Research in Engineering & Technology, 2(3), pp. 107-120, 2014.

[12] D. Khan, “CAKE – Cassifying, associating and knowledge discovery - an approach for distributed data mining (DDM) using parallel data mining agents (PADMAs)”, in WI-IAT '08. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Dec. 9-12, IEEE Xplore Press, Sydney, NSW, pp. 596-601, 2008.

[13] M. Kantardzic, Data mining: concepts, models, methods, and algorithms, 2nd Ed., John Wiley & Sons. New Jersy, 2011.

[14] A. P. Engelbrecht, Computational intelligence: an introduction, 2nd Ed., John Wiley & Sons, Chichester. England, 2007.

[15] D. Driankov, H. Hellendoorn, and M. Reinfrank. An Introduction to Fuzzy Control, 2nd Ed, Springer-Verlag. Berlin, 2013.

[16] R. Heady, G.F. Luger, A. Maccabe and M. Servilla, “The architecture of a Network Level Intrusion Detection System,” Department of Computer Science, College of Engineering, University of New Mexico, 1990, pp. 1-17.

[17] R. Bace and P. Mell, “NIST Special Publication on Intrusion Detection Systems,” Booz-Allen and Hamilton inc, Mclean VA, 2001, pp. 5-22.

[18] R.A. Kemmerer and G. Vigna, “Intrusion Detection : A brief History and Overview,” Computer, 2002 [supplement to security and privacy magazine], pp. 27-30.

[19] Z. Zhou, Y. Xue, J. Liu, W. Zhang, and J. Li. MDH: A High Speed Multi-phase Dynamic Hash String Matching Algorithm for Large-Scale Pattern Set. Information and Communications Security,4861:201– 215, 2007 .

[20] B. Fischer, T. Zoller and J. M. Buhmann, Path based pairwise data clustering with application to texture segmentation, Lecture Notes in Computer Science 2134 (2001) 235–250.

[21] G. Karypis, E.-H. Han and V. Kumar, Chameleon: Hierarchical clustering using dynamic modeling, IEEE Computer (1999) 68–75.

[22] G. Karypis, CLUTO — A Clustering Toolkit , Dept. of Computer Science, University of Minnesota, May 2002. http://www-users.cs.umn.edu/ karypis/cluto/.

[23] S. Zhong and J. Ghosh, A unified framework for modelbased clustering, Journal of Machine Learning Research (2003) 1001–1037.

[24] S. Zhong, T. M. Khoshgoftaar and N. Seliya, Analyzing software measurement data with clustering techniques, IEEE Intelligent Systems (2004) 20–27.

[25] SK Sharma, P Pandey, SK Tiwar “An improved network intrusion detection technique based on k-means clustering via Naïve bayes classification” IEEE Volume 2, Issue 2, February 2012, Issn 2151-961.

[26] M,Varaprsad Rao “Algorithm for Clustering with Intrusion Detection Using Modified and Hashed K – Means Algorithms “Published by IEEE Computer Society,2012.