Towards a Scalable Intrusion Detection System based on Parallel PSO Clustering Using MapReduce

Ibrahim Aljarah* and Simone A. Ludwig
Department of Computer Science
North Dakota State University
Fargo, ND 58102, USA
{ibrahim.aljarah,simone.ludwig}@ndsu.edu

ABSTRACT

The growing data traffic in large networks faces new challenges requiring efficient intrusion detection systems. The analysis of this high volume of data traffic to discover attacks has to be done very quickly. However, in order to be able to process large data, new distributed and parallel methods need to be developed. Several approaches are proposed to build intrusion systems using clustering approaches. In this paper, we introduce an intrusion detection system based on a parallel particle swarm optimization clustering algorithm using the MapReduce framework. The proposed system is scalable in processing large data on commodity hardware.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Clustering; D.2.7 [**Security and Protection**]: Network Security

General Terms

Algorithms, Experimentation, Performance

Keywords

Intrusion detection, Data clustering, Particle swarm optimization, Parallel processing, MapReduce

1. INTRODUCTION

In today's advanced world, there are innumerable network security needs to protect computer systems. One of the most challenging needs is the development of high-accuracy intrusion detection systems (IDSs). This is because of the increasing number of users and the amount of data exchanged which makes it difficult to distinguish the normal data connections from others that contain attacks. This requires the development of IDSs that can analyze large amounts of data patterns in a reasonable amount of time in order to take appropriate actions against the infected patterns.

There are several data mining techniques that have been used for IDSs in the past. Clustering [1] is one of the techniques that is used to explore data. Clustering algorithms find distinct patterns in a group of data without prior knowledge about data labels.

*Corresponding Author

GECCO'13 Companion, July 6–10, 2013, Amsterdam, The Netherlands.
ACM 978-1-4503-1964-5/13/07.

Traditional clustering-based intrusion detection methods suffer from the scalability issue when applied on larger sizes of network traffic and are not easily added to the existing memory resources. Therefore, new parallel scalable methods need to be developed. In general, developing traditional parallel algorithms suffer from a wide range of problems such as inefficient communication, or unfair load balancing, which makes the process of scaling the algorithms to large numbers of processors very difficult.

The MapReduce programming model [2] has become the popular model for parallelizing data-intensive applications due to its features such as fault-tolerance of node failures, and the ability to scale with large numbers of node on commodity hardware.

This paper presents a parallel intrusion detection system (IDS-MRCPSO) based on the MapReduce framework that has been validated as a good parallelization methodology for many applications. In addition, the proposed system incorporates clustering analysis to build the detection model by formulating the intrusion detection problem as an optimization problem.

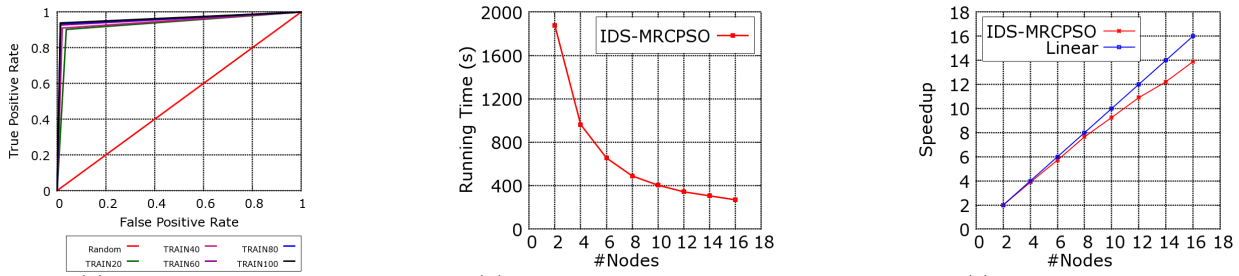
2. PROPOSED INTRUSION DETECTION SYSTEM (IDS-MRCPSO)

Particle Swarm Optimization (PSO) is a swarm intelligence method developed by Kennedy and Eberhart in 1995 [3]. The behavior of PSO mimics the social interaction between individuals such as interactions between the birds in flocks trying to locate an optimal food source. The direction of the movement of each bird is controlled by its current location, the best food location it ever found, and the best food location any bird in the flock ever found.

Exploring large network traffic to detect intrusions takes a long time, thus, our proposed system uses the data clustering concept based on the PSO approach. PSO is parallelized using the MapReduce model as to scale with large-scale network traffic.

The proposed intrusion detection system consists of three main components: preprocessing component, detector model construction component, and validation component. The preprocessing component follows three consequent steps: missing value record elimination, categorical feature elimination, and data normalization.

The detector model construction stage starts by applying the MR-CPSO [4] algorithm to the data results from the preprocessing stage, where only training data is used. The MR-CPSO algorithm has been successfully implemented using



(a) ROC Results (b) TRAIN100 Running Time (c) TRAIN100 Speedup

Figure 1: ROC results, running time and speedup results on the KDD dataset.

the MapReduce programming model, and has shown very good performance and scalability.

MR-CPSO is divided into three main modules: the first module is responsible for updating the particle's centroid vector, the second module is responsible to evaluate the fitness function, and the third module is to merge the outputs of the first and second modules in order to generate a single new swarm.

After the detector model construction stage ends, we extract the global best centroid vector to be used as the detection model in the validation stage. In the validation stage, we used a different record subset called testing dataset to evaluate the detection model by calculating the distances between the testing records and the global best centroids vector (detection model). After that, we assigned the testing records to the closest clusters based on the minimum distances. Finally, the cluster labeling process is triggered to find the correct labels for output clusters generated from the testing record assignment step.

3. EXPERIMENTS AND RESULTS

To evaluate our proposed system, we used the KDD [5] dataset which is a big intrusion detection dataset that has never been fully analyzed by any standard data mining algorithms.

In order to evaluate the impact of the training dataset size in the detector model construction stage, we extracted 5 different samples from the whole training dataset. The sample names consist of the specific format based on the percentage of the whole training dataset. For example, the TRAIN20 sample consists of 20% of the whole training dataset.

We evaluated the IDS's effectiveness by the Receiver Operating Characteristic (ROC) [6] curve which is a plot of the true positive rate against false positive rate. Therefore, we used the Area Under Curve (AUC) measure as the ROC curve evaluation to combine the true positive rate and false positive rate, which is considered a good indicator of their relationship. Figure 1(a) shows the ROC curve using the proposed IDS-MRCPSO system. The figure shows that the best performance and high AUC value are achieved when using TRAIN-100 compared to the other curves. The running times and speedup measures are shown in Figure 1(b) and 1(c), respectively. The speedup results showed reasonable scalability for the proposed system.

4. CONCLUSION

In this paper, we proposed an IDS-MRCPSO system for intrusion detection using the MapReduce methodology to solve the analysis of large-scale network traffic. We have shown that the intrusion detection system can be parallelized efficiently with the MapReduce methodology. The experimental results reveal that IDS-MRCPSO is efficient and scales very close to the optimal speedup by improving the detection results.

Our future plan is to expand the system such as to distinguish between the different types of intrusions, and not only whether an intrusion has occurred or not.

Acknowledgment

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575.

References

[1] J. Han. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, CA, USA, 2005.

[2] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. In *Proceedings of the OSDI '04*, pages 137–150, 2004.

[3] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of the 1995 IEEE International Conference on Neural Networks*, pages 1942–1948. Brisbane, Australia, 1995.

[4] I. Aljarah and S. A. Ludwig. Parallel particle swarm optimization clustering algorithm based on mapreduce methodology. In *Proceedings of NaBIC'12*, pages 104–111, Mexico, November 2012.

[5] S. D. Bay, D. Kibler, M. J. Pazzani, and P. Smyth. The uci kdd archive of large data sets for data mining research and experimentation. *SIGKDD Explor. Newsl.*, 2:81–85, December 2000.

[6] W. Zhu, N. Zeng, and N. Wang. Sensitivity, specificity, accuracy associated confidence interval and roc analysis with practical sas implementations. In *In Proceedings NESUG'10 Conference*, 2010.