Using Genetic Programming for Combining an Ensemble of Local and Global Outlier Algorithms to Detect New Attacks

Gianluigi Folino, Francesco Sergio Pisani, Luigi Pontieri, Pietro Sabatino
(ICAR-CNR, Italy)
name.surname@icar.cnr.it

Maryam Amir Haeri
Amirkabir University of Technology, Iran
haeri@aut.ac.ir

ABSTRACT

Modern intrusion detection systems must be able to discover new types of attacks in real-time. To this aim, automatic or semi-automatic techniques can be used; outlier detection algorithms are particularly apt to this task, as they can work in an unsupervised way. However, due to the different nature and behavior of the attacks, the performance of different outlier detection algorithms varies largely. In this ongoing work, we describe an approach aimed at understanding whether an ensemble of outlier algorithms can be used to detect effectively new types of attacks in intrusion detection systems. In particular, Genetic Programming (GP) is adopted to build the combining function of an ensemble of local and global outlier detection algorithms, which are used to detect different types of attack. Preliminary experiments, conducted on the well-known NSL-KDD dataset, are encouraging and confirm that, depending on the type of attacks, it would be better to use only local or only global detection algorithms and that the GP-based ensemble improves the performance in comparison with commonly used combining functions.

KEYWORDS

Outlier Detection, Ensemble of Classifiers, Intrusion Detection

ACM Reference Format:
Gianluigi Folino, Francesco Sergio Pisani, Luigi Pontieri, Pietro Sabatino and Maryam Amir Haeri. 2019. Using Genetic Programming for Combining an Ensemble of Local and Global Outlier Algorithms to Detect New Attacks. In *Genetic and Evolutionary Computation Conference Companion (GECCO '19 Companion), July 13–17, 2019, Prague, Czech Republic.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3319619.3322018

1 INTRODUCTION

As the number of network connections and the speed of these networks are increasing, the problem of analyzing large streams of data in real time for detecting possible attacks gains relevance in the scientific community of cybersecurity. Typically, *Intrusion Detection Systems* (IDS) are used to detect unauthorized accesses to computer systems and networks (in this case, they are named *Network Intrusion Detection Systems*, NIDS).

Data mining techniques, and, in particular, the ensemble paradigm, have been used mainly for the classification of attacks [3][4], while a few works [7] adopt outlier algorithms to discover anomalies or new types of attack.

On the basis of the analysis of the works in the literature, in this paper, we investigate the feasibility of using an ensemble of outlier algorithms to detect new types of attacks in intrusion detection systems. To this aim, we adopt a genetic programming tool to build the combining function of an ensemble of local and global outlier detection algorithms. In addition, an analysis is conducted in order to assess the performance of local and global outlier algorithms with respect to different types of attack. Preliminary experiments, conducted on the NSL-KDD dataset, confirmed that the combining function built by our approach improves the performance in comparison with commonly used combining functions.

2 METHODOLOGY

Different categorizations of outlier detection algorithms are present in the literature [1]. Here, we consider the categorization that divides the techniques for the detection of the outliers into two categories, on the basis of the kind of outliers the algorithm looks for: global outlier and local outlier algorithms. A global outlier is a data object that significantly deviates from the other samples of the dataset. On the other hand, a local outlier is a data sample that significantly deviates from its neighbors. Combining the scores of different outlier detection by using simple aggregation functions may not obtain good results, as also confirmed by the experimental results. Therefore, we propose to use GP to find a better aggregation function. Indeed, a function in form of a GP tree was used to combine the local and global outlier detection algorithms of the ensemble. As GP engine, the CellulAr GEnetic programming (CAGE) tool [6][5] is adopted. It can run both on distributed-memory parallel computers and on distributed environments and it is based on the fine-grained cellular model. The functions used to build the GP trees are simply the aggregation functions listed in the following: *Avg*, *Max*, *Product*, and *Square*, all of them having as input the outlier scores of some outlier algorithms (both local and global ones) or of other aggregation functions. The functions are replicated with different arity of input, i.e., 3, 4 and 5. The terminal set is constituted by all the outlier detection algorithms considered: i.e., local outlier detection methods LOF (Local Outlier Factor), LoOP (Local Outlier Probabilities), with or without PCA (Principal Component Analysis) for obtaining a 2 dimensionality reduction), and global outlier detection methods SVM (Support Vector Machines) and X-means, with and without PCA.

The fitness function is computed on the basis of a validation set. We suppose that, for a small sample of data, the information

GECCO '19 Companion, July 13–17, 2019, Prague, Czech Republic

ACM ISBN 978-1-4503-6748-6/19/07…$15.00
https://doi.org/10.1145/3319619.3322018

about the real class of the attacks is present. The realistic hypothesis is that a domain expert is able to distinguish attacks and normal connections, at least for a small sample of data. The fitness function is computed as the average classification accuracy of the minority and majority classes over the validation set, which works better than the overall classification accuracy in the case of unbalanced classes [2].

3 EXPERIMENTAL RESULTS

Our main goal is to understand whether local and/or global outlier algorithms can be used to detect new types of attacks and whether GP can be used to develop a combining function of an ensemble of outlier algorithms. The experiments were conducted on the NSL-KDD dataset [9], which was sampled in subsets containing all the normal connections and all the connections belonging to a defined type of attack. No tuning phase was conducted for the GP tool, but the same parameters used in the original paper were used, listed in the following: a probability of crossover equal to 0.7 and of mutation equal to 0.1, a maximum depth equal to 7, a population of 120 individuals and 500 as number of generations. The outlier detection algorithms used for the experiments were taken from the data mining framework ELKI [8]. In more detail, the local algorithms considered contains LOF, LoOP, with a variant using PCA to obtain a two-dimensionality reduction, while global algorithms comprise SVM outlier and *k*-means outlier, the last one employing X-means as clustering algorithm, also considering a variant of the algorithms using PCA.

Experimentally, we discovered that, for the attacks named ipsweep, nmap, and satan, the global outlier techniques work well, while the local outlier techniques obtain good performance against satan, r2l, u2r. With the exception of the attack named portsweep (on which both the techniques work quite well), the other type of technique obtain weak results.

Therefore, we conducted some experiments on using a GP-based approach to combine the scores of the different global and local outlier detection methods, in order to verify whether this ensemble can perform well for all the attacks. A validation set of 5%, 10% or 20% of the entire dataset was used to compute the fitness function of the GP algorithm.

Table 1: AUC value for the ensemble of local and global outlier using the combining function developed by GP for three percentages of validation set (5%, 10% and 20%).

Attack	AvgEnsLoc	AvgEnsGlob	GP Ens. (5%)	GP Ens. (10%)	GP Ens. (20%)
ipsweep	0.43 ± 0.0301	0.67 ± 0.1506	**0.78 ± 0.0812**	0.85 ± 0.0650	0.87 ± 0.0422
nmap	0.37 ± 0.0526	0.66 ± 0.1486	**0.75 ± 0.1400**	0.83 ± 0.1133	0.85 ± 0.0855
portsweep	0.69 ± 0.0596	0.65 ± 0.1101	**0.85 ± 0.2280**	0.88 ± 0.1586	0.90 ± 0.0892
satan	0.63 ± 0.0545	0.58 ± 0.1313	**0.70 ± 0.1653**	0.75 ± 0.0842	0.76 ± 0.0747
r2l	0.78 ± 0.0732	0.59 ± 0.0664	0.69 ± 0.2211	0.74 ± 0.1358	**0.82 ± 0.0958**
u2r	0.81 ± 0.0516	0.40 ± 0.0902	0.67 ± 0.1482	0.75 ± 0.0998	**0.84 ± 0.0772**

Table 1 reports the AUC scores obtained for the different types of attack by using the ensembles of only local and only global algorithms (combined with the average function) and the GP algorithm using three percentages of validation set (5%, 10% and 20%). For each attack, it is reported in bold the minimum percentage of validation set in which GP is significantly better than all the functions both for the global and local ensemble.

It is evident that the GP ensemble, also when only 5% of the tuples is labelled with the real type of attack, obtains considerably good performance in terms of accuracy. For some typologies of attack (ipsweep, nmap, r2l, u2r) using a larger validation set, permits to improve the accuracy. However, for most of the attacks, a 5% validation set is sufficient to outperform both the local and global ensemble. Only for the attacks named r2l and u2r, probably due to the low number of tuples representing these attacks, a 20% validation set produces a significantly better accuracy, even if, by using 5% the validation set, a reasonable level of accuracy is reached. Anyway, as a general consideration, it is necessary to choose a tradeoff between having a small validation set and to obtain a good accuracy.

4 CONCLUSIONS AND FUTURE WORK

A methodology based on an ensemble of local and global outlier detection algorithms is presented. This methodology can be effectively integrated in a general framework using classification, and anomaly detection to detect new types of attack in intrusion detection systems. The novelty of the approach consists in using a module based on an ensemble of outlier detection methods in order to individuate new types of attacks. Genetic programming is used to generate the combining function of the ensemble, which obtains good improvements in terms of AUC in comparison with standard combination functions used for outlier detections. However, the good accuracy obtained by the GP approach is balanced out by the need and the effort required to generate a validation set, necessary to train the GP population. On the contrary, the ensemble of outliers does not require to label a sample of the data, but performs well only for some types of attack.

REFERENCES

[1] Charu C. Aggarwal. 2013. Outlier Ensembles: Position Paper. *SIGKDD Explor. Newsl.* 14, 2 (April 2013), 49–58.

[2] Urvesh Bhowan, Mark Johnston, and Mengjie Zhang. 2012. Developing New Fitness Functions in Genetic Programming for Classification With Unbalanced Data. *IEEE Trans. Systems, Man, and Cybernetics, Part B* 42, 2 (2012), 406–421.

[3] Gianluigi Folino and Francesco Sergio Pisani. 2016. Evolving meta-ensemble of classifiers for handling incomplete and unbalanced datasets in the cyber security domain. *Appl. Soft Comput.* 47 (2016), 179–190.

[4] Gianluigi Folino, Francesco Sergio Pisani, and Pietro Sabatino. 2016. A Distributed Intrusion Detection Framework Based on Evolved Specialized Ensembles of Classifiers. In *Applications of Evolutionary Computation - 19th European Conference, EvoApplications 2016, Porto, Portugal, March 30 - April 1, 2016, Proceedings, Part I (Lecture Notes in Computer Science)*, Vol. 9597. Springer, 315–331.

[5] Gianluigi Folino, Clara Pizzuti, and Giandomenico Spezzano. 2002. Improving Induction Decision Trees with Parallel Genetic Programming. In *10th Euromicro Workshop on Parallel, Distributed and Network-Based Processing (PDP 2002), 9-11 January 2002, Canary Islands, Spain*. 181–18.

[6] G. Folino, C. Pizzuti, and G. Spezzano. 2003. A Scalable Cellular Implementation of Parallel Genetic Programming. *IEEE Transactions on Evolutionary Computation* 7, 1 (February 2003), 37–53.

[7] Aleksandar Lazarevic, Levent Ertoz, Vipin Kumar, Aysel Ozgur, and Jaideep Srivastava. [n. d.]. *A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection.* Chapter 3, 25–36. https://doi.org/10.1137/1.9781611972733.3 arXiv:http://epubs.siam.org/doi/pdf/10.1137/1.9781611972733.3

[8] Erich Schubert, Alexander Koos, Tobias Emrich, Andreas Züfle, Klaus Arthur Schmid, and Arthur Zimek. 2015. A Framework for Clustering Uncertain Data. *PVLDB* 8, 12 (2015), 1976–1987. http://www.vldb.org/pvldb/vol8/p1976-schubert.pdf

[9] M. Tavallaee, E. Bagheri, Wei Lu, and A.A. Ghorbani. 2009. A detailed analysis of the KDD CUP 99 data set. In *Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on*. 1–6. https://doi.org/10.1109/CISDA.2009.5356528