ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Issue 10, October 2016

Augmentation using Cuttlefish Algorithm in Feature Selection for Intrusion Detection Systems

M. Masthan[1], R. Ravi[2]
Research Scholar, Dept. of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli, India[1]
Professor& Head, Dept. of Computer Science and Engineering, Francis Xavier Engineering College, Tirunelveli, India[2]

Abstract:

This paper provides a novel feature selection method based on Cuttle Fish Augmentation (CFA) technique, Intrusion Detection systems (IDSs) are the platforms which will utilize this CFA. One of the serious task of IDSs is to maintain the quality features that represents the entire data, since a large volume of information is handled by the IDSs. The redundant feature were removed by the IDSs. A search scheme proposed in this paper called as Cuttle Fish Augmentation (CFA) which is used to determine the optimal subset features and decision on the particular features will be detected by Decision Tree (DT) classifier, these are created by the CFA. The suggested CFA model is estimated by the KDD Cup 99 data set. The results shows that, the feature subsets acquired by using the CFA produces increased detection rate and increased accuracy rate with a lower false alarm rate,as compared with different features that has been observed earlier and it is also clear that our proposed system is efficient.

Keywords: Feature selection, Cuttlefish algorithm (CFA), Intrusion detection systems (IDSs), Decision trees (DT).

I. INTRODUCTION

As the growth of computer networks, the range of hacking and intrusion incidents is growing 12 months by means of the year as generation rolls out, which has made many researchers focus on building structures referred to as intrusion detection structures (IDSs). Those structures are used to guard computer structures from the threat of theft and intruders (Liao, Lin, Lin, & Tung, 2013). IDSs can be categorized as anomaly detection and misuse detection or signature detection structures (Depren, Topallar, Anarim, & Ciliz, 2005; Wang, Hao, Ma, & Huang, 2010). In anomaly detection, the device builds a profile of that which may be taken into consideration as regular or predicted utilization styles over a time frame and triggers alarms for anything that deviates from this behaviour. Alternatively, in misuse detection, the gadget identifies intrusions based totally on recognized intrusion techniques and triggers alarms through detecting regarded exploits or assaults based totally on their assault signatures. Dimensionality reduction is a generally used step in machine getting to know. Characteristic choice (FS) is a part of dimensional reduction that's called the procedure of selecting a most effective subset of functions that represents the whole dataset. FS has been used in lots of fields, along with classification, data mining, object recognition and so on, and has confirmed to be powerful in getting rid of irrelevant and redundant functions from the original dataset. Given a functional set of size n, the FS problem tries to discover a minimum functional subset of length m (m < n) that enables the development of the first-class classifier with excessive accuracy (Basiri, Ghasem-Aghaee, & Aghdam, 2008). FS has been a fertile discipline of studies and improvement for the reason that 1970s, and it is used efficaciously in the IDSs domain. Stein, Chen, Wu, and Hua (2005) proposed a hybrid genetic-selection tree (DT) version. They used the genetic algorithm (GA) as a generator to supply a premier subset of capabilities, and then the produced functions had been used as an enter for the DT that became built the usage of the C4. five algorithm. Balloon-Canedo, Sanchez-Marono, and AlonsoBetanzos (2011) proposed a brand new combinational technique of discretization, filtering and type which is used as an FS to enhance the type project, and that they applied this method to the KDD Cup 99 dataset. Lin, Ying, Lee, and Lee (2012) offered a smart algorithm which changed into applied to

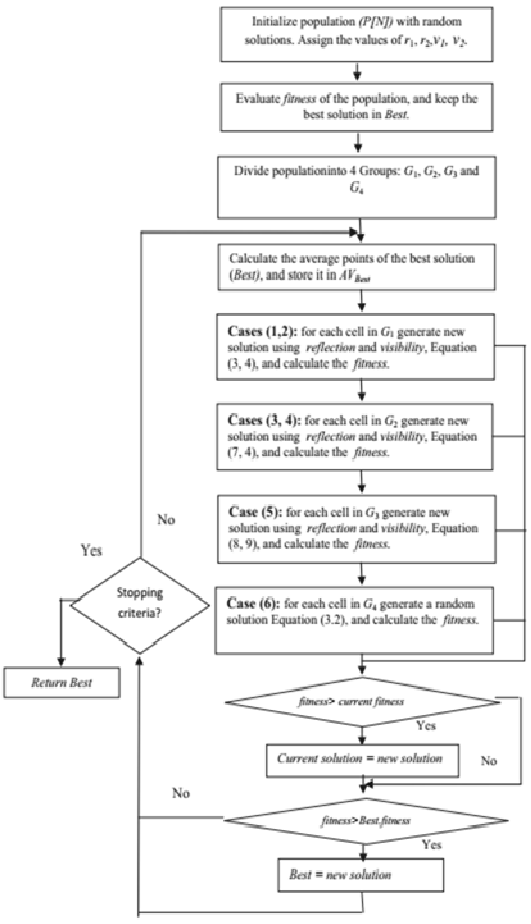
ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Issue 10, October 2016

anomaly intrusion detection. The paper proposed simulated annealing (SA) and aid vector system (SVM) to locate the pleasant function subsets, whilst SA and DT had been proposed to generate decision guidelines to stumble on new assaults. Tsang, Kwong, and Wang (2007) proposed an intrusion detection method to extract correct and interpretable fuzzy IF–THEN policies from network traffic statistics for the category. Inaddition, they used a wrapper genetic FS to produce a superior subset of features. Lasses, Rossi, Sheel, and Mukkamala (2008) proposed a new technique for FS and extraction by means of using the singular cost decomposition paired with the belief of latent semantic analysis, which can find out hidden facts to layout signatures for forensics and in the end realtime IDSs. They used three computerized class algorithms (Maxim, SVM, LGP). Nguyen, Franke, and Petrovic (2010) supplied a time-honoured-characteristic-selection (GeFS) degree to discover worldwide most advantageous feature sets by way of using techniques: the correlation function-selection (CFS) degree and the minimal redundancy-maximal-relevance (murmur) measure. This approach is based totally on fixing a mixed 0–1 linear programming problem by using the use of the department-and-sure set of rules, and the authors carried out the proposed technique to lay out IDSs. A hybrid model based on the information gain ratio and K means is proposed with the aid of Neelakantan, Nagesh, and Tech (2011) to discover 802.eleven-particular intrusions. They used the information advantage ratio as the FS and the ok-manner set of rules because the classifier. Mohanabharathi, Kalaikumaran, and Karthi (2012) proposed a brand new approach which turned into a mixture of the information benefit ratio measure and the ok-means classifier used for FS. As a classifier, they used the returned-propagation set of rules to check these techniques. Considering that IDSs offers with a large quantity of facts, FS is a crucial venture in IDSs. In this paper, we endorse an FS version based totally on the cuttle fish optimization set of rules (CFA) to provide the highest quality subset of features. DT is likewise used as a classifier to enhance the excellence of the produced subsets of features. The relaxation of this paper is organised as follows: segment 2 offers an advent and a brief overview of DT and CFA. The proposed characteristic-choice approach is discussed in segment three. Section four reviews at the experimental consequences of the proposed cuttlefish function-selection technique and a quick dialogue on the obtained effects.

II. CUTTLE FISH ALGORITHM (CFA)

This set of rules mimics the mechanisms in the back of a cuttlefish that are used to change its colour. The patterns and colorations visible in cuttlefish are produced through reflected light from extraordinary layers of cells, such as chromatophores, leucophores and iridophores. The CFA considers two main techniques: reflection and visibility. The reflection manner is used to simulate the mild reflection mechanism, while visibility is used to simulate the visibility of matching styles. These two procedures are used as a search strategy to find the worldwide top-quality answer. The diagram in Fig. 1 of cuttlefish skin, detailing the three main pores and skin systems, two instance, states (a, b) and 3 Fig.1



General Principle of CFA.

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Issue 10, October 2016

awesome ray traces (1, 2, three), indicates the state-of-the-art manner via which cuttlefish can alternate reflective shade (Eric et al., 2012).

CFA reorders these six instances shown in Fig. 1 to be as shown in Fig. 2. The formulation for finding the brand new solution (*newP*) the usage of reflection and visibility is described in Eq. (1):

$$newp¼reflectionþvisibility \quad (1)$$

CFA uses the two processes reflection and visibility to find a new solution. These cases work as a glo- bal search using the value of each point to find a new area around the best solution with a specific interval. The formulations of these processes are described in Eqs. (2) and (3), respectively:

$$reflectionj¼RmG1½i]: Points½j] \quad (2)$$

$$visibilityj¼Vmð Best:Points½j]—G1½i]:Points½j]Þ \quad (3)$$

where, G1 is a set of cells, i is the ith cellular in G1, points[j] represents the jth factor of the ith cellular, great points represents the exceptional answer factors, R represents the degree of reflection, and V represents the visibility diploma of the final view of the pattern. R and V are found as follows:

$$R¼randomðÞmðr1—r2Þ þr2 \quad (4)$$
$$V¼randomðÞmðv1—v2Þ þv2 \quad (5)$$

Wherein, random () feature is used to generate random numbers among (0, 1) and r1, r2, v1, v2 are four consistent values specified by means of the person. As a neighbourhood search, CFA uses cases three and 4 to find the difference between the excellent solution and the present day answer to supply a c program language period across the excellent solution as a brand new search place.

The formula for finding the reflection is as follows:

$$reflectionj¼RmBest: Point½j] \quad (6)$$

While the formulation for finding the visibility remains as in Cases 1 and 2.

The set of rules also uses this situation as a nearby search but this time the difference among the high quality solution factors and the average price of the nice points is used to supply a small place around the best solution as a brand new seeks vicinity. The formulas for finding reflection and visibility in this case are as follows:

$$reflectionj¼RmBest: Points½j] \quad (7)$$
$$visibilityj¼Vmð Best:Points½j]—AVBestÞ \quad (8)$$

Wherein, AVBest is the common price of the exceptional points. Eventually, the CFA uses as the random answers. The overall precept of the CFA is proven.

III. Parametric Analysis

a) Attack Detection Rate

It is the ratio between the total numbers of attack connections detected by our proposed model to the total number of attacks currently available in the data set.

*Attack Detection Rate (ADR) = (Total detected attacks/Total attacks) * 100*

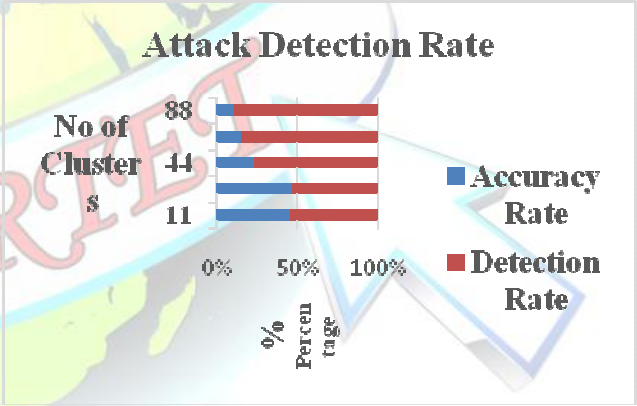


Fig. 2 AR & DR, No. of clusters Vs. Percentage

b) False Positive rate

In information, whilst performing more than one comparison, the time period false wonderful ratio, additionally referred to as the false alarm ratio, commonly refers back to the possibility of falsely rejecting the null speculation for a specific check. The false wonderful charge is calculated as the ratio between the quantity of bad events wrongly labelled as

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Issue 10, October 2016

fine (false positives) and the total range of actual negative activities (regardless of type).

*False Positive Rate (FPR) = (Total misclassified instances/ normal instances) *100*

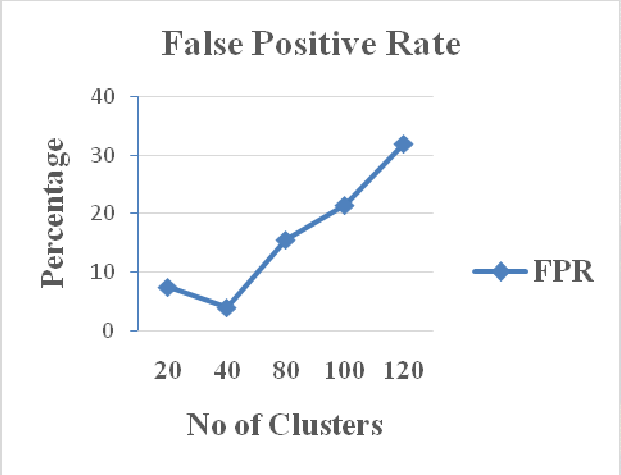


Fig. 3 FPR, Percentage Vs No. of Clusters

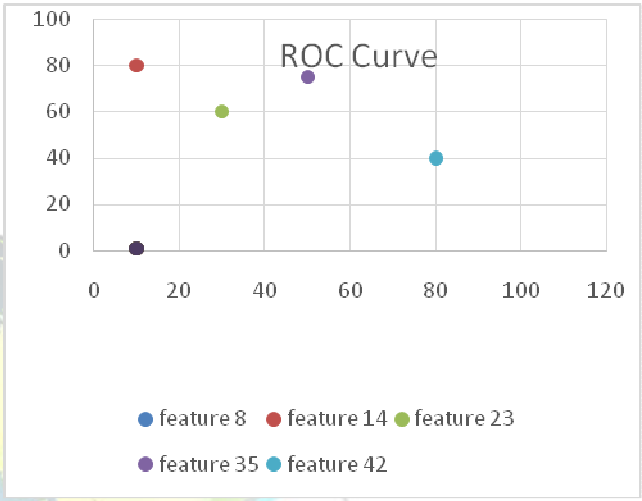


Fig. 4 ROC, Feature Distribution

c) ROC Curve

In records, a receiver running feature curve, or ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is numerous. The curve is created by using plotting the genuine fantastic fee (TPR) in opposition to the false superb rate (FPR) at various threshold settings. The genuine-positive charge is likewise referred to as sensitivity, bear in mind or possibility of detection [1] in gadget gaining knowledge of. The fake-high-quality rate is likewise referred to as the autumn-out or chance of fake alarm [1] and may be calculated as (1 – specificity). The ROC curve is hence the sensitivity as a feature of fall-out. In fashionable, if the possibility distributions for both detection and false alarm are recognised, the ROC curve can be generated through plotting the cumulative distribution characteristic (area beneath the chance distribution from $-\infty$ {displaystyle -infty } -infty to the discrimination threshold) of the detection probability inside the y-axis versus the cumulative distribution function of the false-alarm opportunity in x-axis.

ROC analysis gives gear to select in all likelihood choicest models and to discard suboptimal ones independently from (and previous to specifying) the cost context or the magnificence distribution. ROC analysis is related in an immediate and natural way to price/gain analysis of diagnostic choice making.

IV. Conclusion

In this study, we have investigated the combination model of CFA and DT for feature selection for intrusion detection and evaluated its performance based on the benchmark KDD Cup 99 intrusion data. Firstly, we have modified the CFA to be used as a feature selection tool. Then, we used DT classifier as measurement on the generated features. Empirical results reveal that the produced features are performed the DR and AR especially when the number of produced features was equal or less than 20 features. In general whenever the number of features is decreased, the AR and DR are increased. The value of FPR is not performed during the experiments. It is remained balancing between 3.3 and 3.92. This is because there are some instances of attacks in the test dataset that are never appeared in the train dataset, such as (Mscan, Saint, apache2, mailbomb, processtable, snmpgetattack, snmpguess). The investigation of using CFA as a rule generator for IDS can be suggested as a future work. Moreover, the use of other techniques such as support vector machines, neural networks, clustering methods instead of using DT remains an open issue. Comparisons of feature selection techniques will also provide clues for constructing more effective models for intrusion detection.

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Issue 10, October 2016

REFERENCES

[1] Kml, L., Kittler, J.: Feature set search algorithms. In: Chen, C.H. (ed.) Pattern Recognition and Signal Processing, Sijhoff and Noordhoff, The Netherlands (1978)
[2] Ani, A.A.: An Ant Colony Optimization Based Approach for Feature Selection. In: Proceeding of AIML Conference (2005)
[3] Jensen, R.: Combining rough and fuzzy sets for feature selection. Ph.D. Thesis, University of Edinburgh (2005)
[4] Kohavi, R.: Feature Subset Selection as search with Probabilistic Estimates. In: AAAI Fall Symposium on Relevance (1994)
[5] Kennedy, J., Eberhart, R.C.: Swarm Intelligence. Morgan Kaufmann Publishers Inc., San Francisco (2001)
[6] Dorigo, M., Caro, G.D.: Ant Colony Optimization: A New Meta-heuristic. In: Proceeding of the Congress on Evolutionary Computing (1999)
[7] Bonabeau, E., Dorigo, M., Theraulaz, G.: Swarm Intelligence: From Natural to Artificial Systems. Oxford University Press, New York (1999)
[8] Liu, B., Abbass, H.A., McKay, B.: Classification Rule Discovery with Ant Colony Optimization. IEEE Computational Intelligence 3(1) (2004)
[9] Dorigo, M., Maniezzo, V., Colorni, A.: The Ant System: Optimization by a Colony of Cooperating Agents. IEEE Transactions on Systems, Man, and Cybernetics, Part B 26(1), 29–41 (1996)
[10] Maniezzo, V., Colorni, A.: The Ant System Applied to the Quadratic Assignment Problem. Knowledge and Data Engineering 11(5), 769–778 (1999)
[11] Duda, R.O., Hart, P.E.: Pattern Recognition and Scene Analysis. Wiley, Chichester (1973)
[12] Forman, G.: An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research 3, 1289–1305 (2003)
[13] Pudil, P., Novovicova, J., Kittler, J.: Floating search methods in feature selection. Pattern
Recognition Letters 15, 1119–1125 (1994)
[14] Siedlecki, W., Sklansky, J.: A note on genetic algorithms for large-scale feature selection. Pattern
Recognition Letters 10(5), 335–347 (1989)
[15] Ani, A.A.: Ant Colony Optimization for Feature Subset Selection. Transactions on Engineering,
Computing and Technology 4 (2005)
[16] Zhang, C.K., Hu, H.: Feature Selection Using the Hybrid of Ant Colony Optimization and Mutual Information for the Forecaster. In: Proceedings of the Fourth International Conference on Machine Learning and Cybernetics (2005)
[17] Kanan, H.R., Faez, K., Hosseinzadeh, M.: Face Recognition System Using Ant Colony Optimization-Based Selected Features. In: Proceeding of the First IEEE Symposium on Computational Intelligence in Security and Defense Applications. CISDA 2007, pp. 57–62. IEEE Press, USA (2007) Using ACO-Based Selected Features for Predicting Post-synaptic Activity
[18] Bins, J.: Feature Selection of Huge Feature Sets in the Context of Computer Vision. Ph.D. Dissertation, Computer Science Department, Colorado State University (2000)
[19] Siedlecki, W., Sklansky, J.: On Automatic Feature Selection. International Journal of Pattern Recognition and Artificial Intelligence 2(2), 197–220 (1988)
[20] Dorigo, M., Blum, C.: Ant colony optimization theory: A survey. Theoretical Computer Science 344, 243–278 (2005)
[21] Dorigo, M., Di Caro, G., Gambardella, L.M.: Ant algorithms for discrete optimization. Artificial Life 5, 137–172 (1999)
[22] Engelbrecht, A.P.: Fundamentals of Computational Swarm Intelligence. Wiley, London (2005)
[23] Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishing, Dordrecht (1991)
[24] Pappa, G.L., Baines, A.J., Freitas, A.A.: Predicting post-synaptic activity in proteins with data mining. Bioinformatics 21(2), 19–25 (2005)
[25] Correa, E.S., Freitas, A.A., Johnson, C.G.: A new discrete particle swarm algorithm applied to attribute selection in a bioinformatics dataset. In: The Genetic and Evolutionary Computation Conference - GECCO-2006, Seattle, pp. 35–42 (2006)
[26] Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco (2005)
[27] Shi, Y., Eberhart, R.C.: Parameter selection in particle swarm optimization. In: Porto, V.W., Waagen, D. (eds.) EP 1998. LNCS, vol. 1447, pp. 591–600. Springer, Heidelberg (1998)