Asian Journal of Information Technology 15 (5): 891-895, 2016
ISSN: 1682-3915
© Medwell Journals, 2016

An Effective Intrusion Detection System Using CRF Based Cuttlefish Feature Selection Algorithm and MSVM

[1]K. Rajesh Kambattan and [2]R. Manimegalai
[1]Deptartment of Computer Science and Engineering,
Dhaanish Ahmed College of Engineering, Chennai, India
[2]Deptartment of Computer Science and Engineering,
Park College of Engineering and Ttechnology, Coimbatore, Tamil Nadu, India

Abstract: In this study, we propose an effective intrusion detection system for improving the detection accuracy. In this proposed system, we propose a new feature selection algorithm called Enhanced Cuttlefish Feature Selection Algorithm (ECFSA) for effective feature selection and Intelligent Agent based Enhanced Multiclass Support Vector Machine (IAEMSVM) classification algorithm is used for classification. The experimental results of the proposed system show that this system produced high-detection rate when tested with KDD Cup 99 data set.

Key words: Cuttle fish, multiclass support vector machine, intrusion detection system, feature selection, agent

INTRODUCTION

An Intrusion Detection System (IDS) is playing a major role to provide the security through monitoring service and to identify the malicious behavioral users in the networks. It is alsoreports to the administrator of the system/network in charges through messages or alarm signal. Generally, this system can be categories into two types such as network based IDS (Network level identification) and Host based IDS (System level identification). Malicious activities are called as attacks, these are grouping into two categories such as passive and active attacks. Passive attacks mean the taking privilege into the system/network and gains knowledge but not affecting others. Active attacks, gains knowledge from network/system with affecting or damaging others. Another important concept in IDS is alert, these alerts are worked based on four scenario's. There are, True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN). From these scenarios, second and third must be handled carefully. Because of these are capable to lead our network or system into big damage.

Attribute selection is also called as feature selection which is used for selecting or retrievingrelevant features for solving the problem. The main advantage of the feature selection is to save time and improve the classification accuracy. In network data set (KDD CUP' 99) consists of huge volume of data, from this dataset the system selects only relevant attributes and also applied for better result. It is used to reduce the computation time and improve the performance of the system. Feature selection selects only the suitable features according the problem. This feature selection and classification methods are the best choices for identifying the attacks in computer networks to make better result.

Classification is one of the major technique in data mining, it is categories the items of given dataset as two classes/groups. The main advantage of classification is to segregate the given items from huge data. For example, a classification algorithm is used to identify items (eatable items/non eatable items, etc. which are stored in separate places in supermarket for easy access by people. In IDS, classification algorithm is classified into two classes such as normal and abnormal. Generally, the classification start with simple model like binary classification that should be classify only two categories, i.e., eatable or non-eatable items. At the same time, this classification also can be expanded into deeper level based on problems to categories the records into many classes. Classification accuracy is calculated based on the difference between the prediction and actual value.

In this study, we propose an effective intrusion detection system for improving the detection accuracy. The proposed system is the combination of newly proposed feature selection algorithm is called Enhanced

Corresponding Author: Rajesh Kambattan, Deptartment of Computer Science and Engineering, Dhaanish Ahmed College of Engineering, Chennai, India

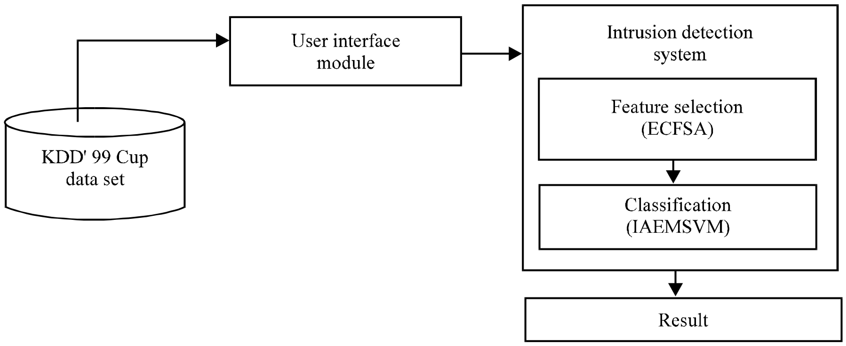

Fig. 1: System architecture

Cuttlefish Feature Selection Algorithm (ECFSA) and Intelligent Agent based Enhanced Multiclass Support Machine (IAEMSVM) for making effective decision over the dataset.

Literature review: Many works have been done in this direction in the past by various researchers. Among them, Laurentys *et al.* (2011) proposed a novel approach for detecting errors in immune system. Their approach shows that better detection rate and false alarm. Nadiammai and Hemalatha (2014) proposed IDS with the combination of data mining technique to identify the relevantinterest data for the user to save time. The proposed EDAT algorithm solves classification problem, labeled data and distributed DDoS. Altwaijry and Algarny (2012) proposed a Bayesian probability based IDS for effective classification. This classification algorithm is trained by Apriori using data set 99 and detects the intrusion in superior nature.

Barani (2014) proposed model based Genetic algorithm and AIS called GAIS which is going to be applied in dynamic environment MANET. The GAIS detects topological changes in two styles, partial and total. Some kind of spherical detector is used to identify anomaly in network traffic. Shakshuki *et al.* (2013) presented a secure IDS model for MANET, it belongs to EAACK (Enhanced Adaptive Acknowledgment). It guaranties that highest malicious behavior detection rates when comparing with others. Credit scoring imbalance can be solved by classification algorithms proposed by Brown and Mues (2012). Random forest and gradient booster classification algorithms are used to produce best credit scoring. Al Snousy *et al.* (2011) proposed supervised machine learning model through classification mechanism. Here many classification mechanisms are compared but classification accuracy can be calculated by gain ratio attributes selection method. Gene's differences in cancer cells are easily explored by gain ratio method.

A neuro fuzzy classification technique was proposed by Ganapathy *et al.* (2014). The researchers suggested that inputs given basis on fuzzified format. Because of the different kinds of data are easily identified by machine learning algorithms. Here, the proposed method is to more powerful than RBFNN and ANFIS algorithms. Juhola *et al.* (2014) proposed data cleaning process using neighborhood method to achieve classification accuracy. In distributed data environment data should be scattered, the classifying process leads to headache for further process to avoid such kind of complication use neighborhood cleaning method. Abdelhamid (2015) published multi label classification concept with association of associative extraction method which is called as EMCAC (Enhanced Multi-label Classifiers based Associative Classification), it avoids the problem to produce largest frequency class and discard all other relevant data. This study helps to users to avoid phishing data related problems and make correct decision. Graph constructed from original feature space is not advisable idea which leads to noisy and non-linear data distribution may occur. Wang *et al.* (2015) proposed two graph regularized NMF (Nonnegative Matrix Factorization) methods such as $AGNMF_{FS}$ and $AGNMF_{MK}$. The proposed methods adapt feature selection and kernel learning to produce better result. Ganapathy *et al.* (2012) proposed an effective intrusion system by the help of intelligent agents. This system has been improved by the uses of effective rules by same researcher in later on (Ghosh *et al.*, 2014).

System architecture: The overall system architecture of the proposed system is shown in Fig. 1. The proposed system architecture is consists of four major components such as KDD'99 cup data set, user interface module, intrusion detection system and result.

The user interface module collects the necessary data from dataset and sent it to the intrusion detection system for further processing. The intrusion detection system is consists of two components such as feature selection and classification. This feature selection component is used an effective proposed cuttlefish algorithm for optimizing/selecting the features. The classification component is used an Intelligent Agent based Enhanced Multiclass Support Vector Machine (IAEMSVM). Features have been selected by feature selection component of the intrusion detection system and also it classifies the records as normal or abnormal by using the classification component with the help of Enhanced Cuttlefish Feature Selection algorithm. Finally, it sends the classified records into the result module.

MATERIALS AND METHODS

Proposed work: This study discusses about the proposed system. In this study, we have discussed in detail about the proposed Enhanced Cuttlefish Feature Selection Algorithm (ECFSA) and also explain the role of IAEMSVM algorithm in this proposed system.

Feature selection: The proposed feature selection algorithm is proposed according to the existing cuttlefish algorithm (Eesa *et al*., 2015) for optimal feature selection. The proposed feature selection algorithm is changing the decision over the feature selection process based on the environments. In this algorithm, two different processes such as reflection and visibility have been applied for finding necessary features in each subset. Five case studies also have been carried out using the different searching techniques like local and global search. Finally, the best subset of features has been selected from best subset and average best subset.

In this study, we propose anew feature selection algorithm called Enhanced Cuttlefish Feature Selection algorithm according to ICRF (Ganapathy *et al*., 2016) and CRF (Gupta *et al*., 2010) for effective feature selection over the intrusion datasets. Here, selects the best features by applying the multistage evaluation process based on the removal of worst features from subsets in every stage.

The ECFSA reorders all the possible cases which are used in existing cuttlefish algorithm. The formulation for finding the New Solution (NS) using Reflection (R) and Visibility (V) is described in Eq. 1:

$$\mathrm{NS} = \mathrm{R} + \mathrm{V} \tag{1}$$

The ECFSA uses the two processes reflection and visibility to find a new solution. These cases work as a global 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 Eq. 2 and 3, respectively:

$$\mathrm{R_j = DRC_1[i].Points\,[j]} \tag{2}$$

$$\mathrm{V_i = VD\left(B.Points[j] - C_i[i]\ Points[j]\right)} \tag{3}$$

Where:
C_1 = A group of cells,
I = The ith cell in C_1
Points [j] = The jth point of the ith cell
Best (B). Points = The best solution points
DR = The degree of reflection
VD = The visibility degree of the final view of the pattern

The DR and VD are found as follows:

$$\mathrm{DR = ICFR \times (dr_1 - dr_2) + dr_2} \tag{4}$$

$$\mathrm{VD = ICRF \times (vd_1 - vd_2) + vd_2} \tag{5}$$

Where:
ICRF = The function is used to generate random numbers between (0,1) based on conditions
dr_1, dr_2, vd_1, vd_2 = The four constant values specified by the user

As a local search, ICRFCFA (Gupta *et al*., 2010) used to find the difference between the best solution and the current solution for producing an interval around the best solution as a new search area in this research. The formula for finding the reflection is as follows:

$$\mathrm{R_j = DR \times B.Point[j]} \tag{6}$$

While the formulation for finding the visibility remains. The proposed algorithm also uses this case as a local search, but this time the difference between the best solution points and the average value of the Best (B) points is used to produce a small area around the best solution as a new search area. The formulas for finding reflection and visibility in this case are as follows:

$$\mathrm{R_j = DR \times B.Point[j]} \tag{7}$$

$$\mathrm{V_j = VD \times (B.Points[j] - AV_B)} \tag{8}$$

Where, AV_B is the average value of the Best points. Finally, the ICRFCFA provides the selected features for effective decision making.

Enhanced cuttlefish feature selection algorithm: Input: KDD'99 cup dataset; output: selected features

- Step 1: initialize the population (features) with random subset
- Step 2: Evaluate fitness of the population using EMSVM
- Step 3: Store the best subset in B
- Step 4: Remove one feature from B using ICRF (Ganapathy *et al*., 2016)
- Step 5: Sort the original features in descending order based on the fitness value which is calculated according to (Gupta *et al*., 2010)
- Step 6: Randomly selected features is split into two and store into a set
- Step 7: Find the Reflection subset from randomly selected set using ICRF
- Step 8: Find the Visibility set for removing the elements of R using ICRF
- Step 9: New subset is created by using the features of visibility and reflection
- Step 10: Evaluate the new subset using EMSVM
- Step 11: If the new subset is better than the set B then the current new subset is considered as B

The proposed feature selection algorithm selects the best subset from resulted features which are finalized by the evaluation process using fitness function. Ascending the feature subsets and selects features using Intelligent CRF in different form such as reflection and visibility. This algorithm is useful for selecting optimal features which are useful for making effective decision over the dataset.

Classification: We have used the existing classification algorithm called Intelligent Agent based Multiclass Support Vector Machine (IAEMSVM) algorithm for effective classification (Ganapathy *et al*., 2012). This algorithm uses the clustering technique, decision tree and intelligent agent for improving the classification accuracy.

RESULTS AND DISCUSSION

KDD'99 Cup data set: The KDD'99 Cup dataset is used for carrying out the experiments. This dataset contains 5 million records and each connection record is described by 41features. It has 22 categories of attacks fromthe following four classes such as DoS, R2L, U2R and Probe. It has 391458 DOS attack records, 52 U2R attack records, 4107 probe attack record, 1126 R2Lattack records and 97278 normal records only in this 10percent of this dataset.

Experimental results: Five experiments have been conducted with different number of records. We have

Table 1: Performance of the proposed intrusion detection system

Ex. No.	ECFSA (10 features) +IAEMSVM	ECFSA (15 features) +IAEMSVM	ECFSA (20 features) +IAEMSVM
1	99.47	99.37	99.31
2	99.38	99.29	99.24
3	99.28	99.21	99.14
4	99.35	99.23	99.11
5	99.32	99.24	99.13

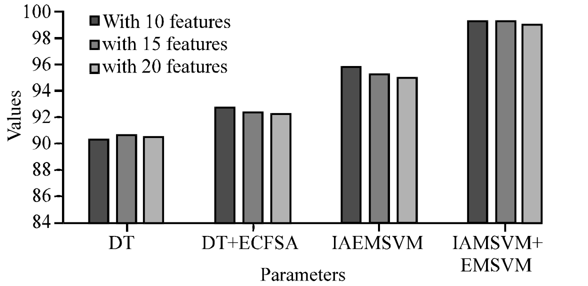

Fig. 2: Comparative analysis

Experimental results: Five experiments have been conducted with different number of records. We have used the three different best features set which are selected by ECFSA. These feature sets contains 10, 15 and 20 features. Table 1 shows the most important 10 features which are selected by the proposed feature selection algorithm from 41 features of KDD'99 Cup dataset.

List of most 10 important selected features:

- Duration
- Protocol
- Type
- Service
- Flag
- Src_bytes
- Hot
- Num_failed_logins
- Num_access_files
- Dst_host_srv_serror_rate
- Dst_host_rerror_rate

It can be observed that the performance of the proposed system is better with 10 features selected by ECFSA. The reason for this different accuracy produced by the proposed system is due to the uses of different number of features selected by ECFSA. According to the ECFSA, can be considered the best accuracy produced by the system with the use of particular subset. So, 10 features contained feature set is a best feature set.

Figure 2 shows the comparative analysis between the proposed system (ECFSA+IAEMSVM) and the existing systems. Here, we have considered three different sets which are contain 10, 15 and 20 features.

From this Fig. 2, it can be observed that the performance of the proposed system is better than the existing system when we have considered all three different numbers of selected features for the experimental analysis. The reason for the better performance of the proposed system is the uses of effective classification algorithm.

CONCLUSION

An effective intrusion detection system is proposed and implemented in this paper for improving the detection accuracy. The proposed Enhanced Cuttlefish Feature Selection Algorithm (ECFSA) selects the necessary features and sends it for classifying the data to Intelligent Agent based Enhanced MSVM (IAEMSVM). The performance of the feature selection algorithm has been validated by the use of an effective classification algorithm. Future researches in this direction could be the use of fuzzy rules for selecting features and classification.

REFERENCES

Abdelhamid, N., 2015. Multi-label rules for phishing classification. Applied Comput. Inf., 11: 29-46.

Al Snousy, M.B., H.M. El-Deeb, K. Badran and I.A. Al Khlil, 2011. Suite of decision tree-based classification algorithms on cancer gene expression data. Egypt. Inf. J., 12: 73-82.

Altwaijry, H. and S. Algarny, 2012. Bayesian based intrusion detection system. J. King Saud Univ. Comput. Inf. Sci., 24: 1-6.

Barani, F., 2014. A hybrid approach for dynamic intrusion detection in ad hoc networks using genetic algorithm and artificial immune system. Proceedings of the 2014 Iranian Conference on Intelligent Systems, February 4-6, 2014, Bam, pp: 1-6.

Brown, I. and C. Mues, 2012. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Syst. Applic., 39: 3446-3453.

Eesa, A.S., Z. Orman and A.M.A. Brifcani, 2015. A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. Expert Syst. Applic., 42: 2670-2679.

Ganapathy, S., P. Yogesh and A. Kannan, 2012. Intelligent agent-based intrusion detection system using enhanced multiclass SVM. Comput. Intell. Neurosci. 10.1155/2012/850259.

Ganapathy, S., K. Kulothungan, S. Muthurajkumar, M. Vijayalakshmi, P. Yogesh and A. Kannan, 2013. Intelligent feature selection and classification techniques for intrusion detection in networks: A survey. EURASIP J. Wireless Commun. Network. 10.1186/1687-1499-2013-271

Ganapathy, S., P. Vijayakumar, P. Yogesh and A. Kannan, 2016. An intelligent CRF based feature selection for effective intrusion detection. Int. Arab J. Inf. Technol., 10: 44-50.

Ghosh, S., S. Biswas, D. Sarkar and P.P. Sarkar, 2014. A novel neuro-fuzzy classification technique for data mining. Egypt. Inf. J., 15: 129-147.

Gupta, K.K., B. Nath and R. Kotagiri, 2010. Layered approach using conditional random fields for intrusion detection. IEEE Trans. Dependable Secure Comput., 7: 35-49.

Juhola, M., H. Joutsijoki, H. Aalto and T.P. Hirvonen, 2014. On classification in the case of a medical data set with a complicated distribution. Applied Comput. Inf., 10: 52-67.

Laurentys, C.A., R.M. Palhares and W.M. Caminhas, 2011. A novel artificial immune system for fault behavior detection. Expert Syst. Applic., 38: 6957-6966.

Nadiammai, G.V. and M. Hemalatha, 2014. Effective approach toward intrusion detection system using data mining techniques. Egypt. Inf. J., 15: 37-50.

Shakshuki, E.M., N. Kang and T.R. Sheltami, 2013. EAACK-A secure intrusion-detection system for MANETs. IEEE Trans. Ind. Electron., 60: 1089-1098.

Wang, J.J.Y., J.Z. Huang, Y. Sun and X. Gao, 2015. Feature selection and multi-kernel learning for adaptive graph regularized nonnegative matrix factorization. Expert Syst. Applic., 42: 1278-1286.