International Journal of Electrical and Computer Engineering (IJECE)
Vol. 6, No. 5, October 2016, pp. 2432 – 2436
ISSN: 2088-8708

Review of IDS Develepment Methods in Machine Learning

Abdulla Amin Aburomman* **and Mamun Bin Ibne Reaz***
*Department of Electrical, Electronic & Systems Engineering, Faculty of Engineering & Built Environment, National University of Malaysia, 43600 UKM Bangi, Selangor, Malaysia

Article Info

Article history:
Received May 24, 2016
Revised Jul 10, 2016
Accepted Jul 25, 2016

Keyword:
Clustering
Ensemble methods
Hybrid system
IDS
Machine learning

ABSTRACT

Due to the rapid advancement of knowledge and technologies, the problem of decision making is getting more sophisticated to address, therefore the inventing of new methods to solve it is very important. One of the promising directions in machine learning and data mining is classifier combination. The popularity of this approach is confirmed by the still growing number of publications. This review paper focuses mainly on classifier combination known also as combined classifier, multiple classifier systems, or classifier ensemble. Eventually, recommendations and suggestions have also included.

Corresponding Author:
Abdulla Amin Aburomman
Department of Electrical, Electronic & Systems Engineering, Faculty of Engineering & Built Environment, National University of Malaysia
43600 UKM Bangi, Selangor, Malaysia.
Email: reoroman@hotmail.com

1. INTRODUCTION

In today's huge on-line communications, safeguarding the precious information from slipping into the hands of hackers is the greatest obstacle. In spite of these types of risks, the IDS try very hard to fight the cyber-attacks. IDS is sorted to misuse and anomaly detection. In misuse detection, the IDS evaluate the data it collects and compares it to the huge data source of attack signatures that define various attack kinds. In anomaly detection, the system administrator identifies the normal state of network's traffic, and then any identification of pattern which does not conform to an anticipated saved normal state will be identified anomaly. IDS can be seen as pattern recognition. There are three methods of pattern recognition, (i) data acquisition, where data are gathered. (ii) Data processing, where data are processed to eliminate redundant features, and (iii) pattern classification. There are some challenges in pattern classification. First, the huge volume of data; second, finding an effective technique to cope using numeric features; finally, research in the area of pattern recognition show that binomial distributions cannot represent its behavior, meaning that conventional methods of parametric statistical might not assist. Finally, pattern recognition issues involve other kinds of classification including intrusion detection. There are several well-known datasets used in the analysis of IDS. KDD cup 99 dataset is most famous one, followed by NSL-KDD which is recommended to solve a number of the inherent issues in KDD'99.

2. RELATED WORK

Different approaches have implemented to create a perfect IDS using data mining and machine learning methods.

Patel and Buddhadev [1], proposed an architecture of hybrid IDS based on misuse and anomaly detection. They used Snort software (free and open source software for IDS and IPS) to capture and analyze network packets. They used string searching algorithm called "AhoCorasick algorithm" to compare the incoming pattern with saved one in the signature database, if there is a match, an alarm will rise, if not, the pattern will be passed to anomaly detector for further classification. Yet, the authors did not describe which algorithm they used in the anomaly model, nor provide experiments based on their suggested model.

Hlaing, Thuzar [2], proposed feature selection based on Mutual Correlation method to reduce the 34 continuous KDD 99 dataset features to 10. He utilized Fuzzy Decision Tree as a classifier to differentiate between normal and 4 classes of attack. He compares his approach with Neural Network+ SVM, Fuzzy Logi, and C4.5. The author proves that his approach could compete others in term of accuracy, though it could be great in terms of comparison if the author implemented the Mutual Correlation feature selection with other classifiers as well, especially with the strong C4.5 DT classifier.

Chandrashekhar and Raghuveer [3] evaluates 4 clustering methods: fuzzy c-means, Mountai, Subtractive, and k-means clustering using the well known KDD 99 dataset. Their results show that fuzzy c-means and k-means clustering performed better in terms of computation time and accuracy.

Taghanaki et.al [4], combined two feature extraction methods, LDA, and PCA based on RBF Neural Network as pattern classifier. Utilizing Weka (Data Mining software), they used KDD 99 dataset for evaluating their approach and compare the results against Kernel Discriminant Analysis (KDA), Local Liner Embedding (LLE), Principal Component Analysis (PCA), and Linear Discriminant Analysis (LDA). Their experiments indicates that their proposed approach could achieve better results.

Yingmei and Songtao [5] proposed classification in ad hoc networks based on improved k-means clustering algorithm and Hybrid Genetic Algorithm (HGA). The improved k-means clustering used to split the data to normal and anomaly traffic, and the HGA used to classify the intrusion behavior. Using KDD 99 for the experiment, the results show improved detection accuracy and low false positive (FP) rate.

3. SINGLE PATTERN RECOGNITION

Earlier times, pattern recognition concentrated on developing single classifiers. The Vast majority of these approaches are well recognized among pattern recognition and machine learning communities. The following is a brief history about well know single classifiers.

- Fuzzy logic: it is a potential technique suggested by Zadeh (1965), to cope with decision-making strategies by applying IF-THEN rules. It can solve the non-linear problems and can provide a linguistic representation. Liu et.al. [5], proposed IDS model based on fuzzy logic and (Naïve Bayes (NB) classifiers, where fuzzy system employed to evaluate the potential threats. The results show that fuzzy system could decrease the false alarm rate and provide better evaluation of the potential threats.
- Artificial Neural Networks (ANN): it is one of the most current effective classification methods. Versatility and the natural speed are the advantages of choosing ANN in the data classification. It can handle the multi-variables, non-linear data sets. Bitter et.al. [6], discussed critical cases in intrusions like spam, worm, and DoS being resolved by ANN. He reports that dataset characteristics, such as size, format, and dimensionality are very critical in order to model a successful ANN.
- K-Nearest Neighbors: it is well-known classification algorithm, which utilizes distance measurement. It considers that the whole selection of sample consists of the perfect classification for each and every single item. To classify a new object, the algorithm calculates the distance between every object and considers objects that are near to each other are from the same class.
- Support Vector Machine (SVM): is a technique created by Vapnik (1998). SVM construct a hyperplane between two datasets and try to maximize the margin between two classes to improve classification accuracy.
- Naïve Baye (NB): broadly utilized method in classifications purposes. It assumes that each feature has its own independency among others. It is based on Directed Acyclic Graph (DAG), where nodes are used to depict the features and arcs depict their dependencies.
- Decision Trees (DT): In DT classification the feature attitudes explaining more details about the information. For an efficient classification, the features with highest information gain (IG) are the better. DT contains nodes, arcs (edges), and leaves. Nodes represent the segmented features, arcs (edges) is the outcome of any node (children of that node), and leaves represent the classified class using a decision value.

4. HYBRID AND ENSEMBLE PATTERN RECOGNITION

The hybrid and ensemble classification methods seek to combine more than one classifier to boost their efficiency in order to improve the classification accuracy and help to understand different problems. In literature, several approaches for classifiers combination proposed.Table 1 illustrates the detailed numbers of the articles used hybrid and ensemble methods.

Table 1. Homogeneous ensembles for IDS

Hybrid Classifiers	Ensemble Classifiers
[7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21],	[22, 23, 24, 25, 26, 27, 28, 29, 30, 31]

5. DISCUSSION AND CONCLUSION

The above approaches lead to the subsequent issues:

- Data hybridization and knowledge related issues.
    1. Explicitly and constancy in knowledge and data.
    2. Privacy of data.
    3. Integration between knowledge and data.
    4. Cost of data acquire.
- Classification issues in the hybrid system.
    1. Taking into consideration the diversity between classifiers ensemble, and processing time.
    2. Utilize voting strategy in the ensemble.
    3. Utilize other functions, such as parametric model

The quality of designing classifier depends on a good prior knowledge. If the learning was incomplete or unrepresentative, this may create a sub-standard classifier. It is very useful to not employ the data from the same source. Besides, subsequent questions also should be satisfy:

1. Does combining data taken from undependable resources going to reduce classification quality? and what is the quality of such data?
2. How to combine different classifires. i.e. we can train different classifiers on different subset of data, then we deside which method to use to combine them, still there are probelms regarding the qulity method of learning.
3. Is the classifier learning on consistent material? If we would like to combine another materials for learning the classfier taken from other source, then such combinations could produce instability.

   Besides, instability classification methods should be analyzed the following:

   - Instability classification methods and removal out of the actual rule set.
   - Instability classification methods in other learning data set.
   - Instability classification methods and removal between learning samples and rules.
4. How to satisfy limits enforced on data source? it is generally under resitrection of law due to privacy reasons. so we should take into account the safety of privacy.
5. Nowdays, making decision with high-quality could be in hand, but vey expensive. This is a cost-sensitive information relation issue. i.e. the trade-off between data cost and expected medical diagnosis results in medical scenario.

We observed the above issues. We also observed that many studies did not consider classifier combination based on feature space partitioning, hybrid classifiers based on one-class classification paradigm, or classifier ensemble for data stream classification. This should be good motivation for future research.

REFERENCES

[1] K. K. Patel and B. V. Buddhadev, "An architecture of hybrid intrusion detection system," *International Journal of Information and Network Security*, vol. 2, no. 2, p. 197, 2013.

[2] T. Hlaing, "Feature selection and fuzzy decision tree for network intrusion detection," *International Journal of Informatics and Communication Technology (IJ-ICT)*, vol. 1, no. 2, pp. 109–118, 2012.

[3] A. Chandrashekhar and K. Raghuveer, "Performance evaluation of data clustering techniques using kdd cup-99 intrusion detection data set," *International journal of information and network security*, vol. 1, no. 4, p. 294, 2012.

[4] S. A. Taghanaki, B. Z. Dehkordi, A. Hatam, and B. Bahraminejad, "Synthetic feature transformation with rbf neural network to improve the intrusion detection system accuracy and decrease computational costs," *International Journal of Information and Network Security*, vol. 1, no. 1, p. 28, 2012.

[5] L. Liu, P. Wan, Y. Wang, and S. Liu, "Clustering and hybrid genetic algorithm based intrusion detection strategy," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 12, no. 1, pp. 762–770, 2014.

[6] C. Bitter, D. A. Elizondo, and T. Watson, "Application of artificial neural networks and related techniques to intrusion detection," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2010, pp. 1–8.

[7] B. Shanmugam and N. B. Idris, *Hybrid intrusion detection systems (HIDS) using Fuzzy logic*. INTECH Open Access Publisher, 2011.

[8] A. Herrero and E. Corchado, *Mobile Hybrid Intrusion Detection*. Springer, 2014.

[9] R. Mandal and S. Yadav, "An improved intrusion system design using hybrid classification technique," *International Journal of Computer Applications*, vol. 117, no. 10, 2015.

[10] H. Bostani and M. Sheikhan, "Hybrid of binary gravitational search algorithm and mutual information for feature selection in intrusion detection systems," *Soft Computing*, pp. 1–18, 2015.

[11] R. Sujendran and M. Arunachalam, "Hybrid fuzzy adaptive wiener filtering with optimization for intrusion detection," *ETRI Journal*, vol. 37, no. 3, pp. 502–511, 2015.

[12] A. Dhivya and S. Sivanandan, "Hybrid fuzzy jordan network for robust and efficient intrusion detection system," *Indian Journal of Science and Technology*, vol. 8, no. 34, 2015.

[13] S. Mourougan and M. Aramudhan, "Hybrid evolutionary algorithmn based intrusion detection system for denial of service attacks," *Indian Journal of Science and Technology*, vol. 8, no. 35, 2015.

[14] S. Dubey and J. Dubey, "Kbb: A hybrid method for intrusion detection," in *Computer, Communication and Control (IC4), 2015 International Conference on*. IEEE, 2015, pp. 1–6.

[15] Y. Canbay and S. Sagiroglu, "A hybrid method for intrusion detection," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2015, pp. 156–161.

[16] S. K. Sharma, D. Bhattacharyya, M. R. Patra, and T.-h. Kim, "A new parallel hybrid model-intrusion prevention systems," in *2015 8th International Conference on Security Technology (SecTech)*. IEEE, 2015, pp. 17–24.

[17] T. Patil and B. Joshi, "Improved acknowledgement intrusion detection system in manets using hybrid cryptographic technique," in *2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*. IEEE, 2015, pp. 636–641.

[18] G. P. Rout and S. N. Mohanty, "A hybrid approach for network intrusion detection," in *Communication Systems and Network Technologies (CSNT), 2015 Fifth International Conference on*. IEEE, 2015, pp. 614–617.

[19] M. E. Haque and T. M. Alkharobi, "Adaptive hybrid model for network intrusion detection and comparison among machine learning algorithms," *International Journal of Machine Learning and Computing*, vol. 5, no. 1, p. 17, 2015.

[20] K. Kaur and N. Kaur, "A hybrid approach of fuzzy c-mean clustering and genetic algorithm (ga) to improve intrusion detection rate," *International Journal of Science and Research*, 2015.

[21] A. Tesfahun and D. L. Bhaskari, "Effective hybrid intrusion detection system: A layered approach," *International Journal of Computer Network and Information Security*, vol. 7, no. 3, p. 35, 2015.

[22] L.-W. Chen, "Network intrusion detection model with clustering ensemble method," *International Journal of Security and Its Applications*, vol. 9, no. 11, pp. 239–250, 2015.

[23] A. Cuzzocrea, G. Folino, and P. Sabatino, "A distributed framework for supporting adaptive ensemble-based intrusion detection," in *Big Data (Big Data), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1910–1916.

[24] M. Sreenath and J. Udhayan, "Intrusion detection system using bagging ensemble selection," in *Engineering and Technology (ICETECH), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1–4.

[25] M. Milliken, Y. Bi, L. Galway, and G. Hawe, "Ensemble learning utilising feature pairings for intrusion detection," in *2015 World Congress on Internet Security (WorldCIS)*. IEEE, 2015, pp. 24–31.

[26] P. Sornsuwit and S. Jaiyen, "Intrusion detection model based on ensemble learning for u2r and r2l attacks," in *2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE)*. IEEE, 2015, pp. 354–359.
[27] D. Gaikwad and R. C. Thool, "Intrusion detection system using bagging ensemble method of machine learning," in *Computing Communication Control and Automation (ICCUBEA), 2015 International Conference on*. IEEE, 2015, pp. 291–295.
[28] P. Amudha, S. Karthik, and S. Sivakumari, "Intrusion detection based on core vector machine and ensemble classification methods," in *Soft-Computing and Networks Security (ICSNS), 2015 International Conference on*. IEEE, 2015, pp. 1–5.
[29] L. Nan and X. Chun-Zhi, "A wavelet transform based support vector machine ensemble algorithm and its application in network intrusion detection," in *Intelligent Systems Design and Engineering Applications (ISDEA), 2014 Fifth International Conference on, vol*, vol. 109, 2014, pp. 15–16.
[30] B. A. Tama and K. H. Rhee, "A combination of pso-based feature selection and tree-based classifiers ensemble for intrusion detection systems," in *Advances in Computer Science and Ubiquitous Computing*. Springer, 2015, pp. 489–495.
[31] A. A. Aburomman and M. B. I. Reaz, "A novel SVM-kNN-PSO ensemble method for intrusion detection system," *Applied Soft Computing*, vol. 38, pp. 360–372, 2016.