International Journal of Advances in Applied Sciences (IJAAS)
Vol. 7, No. 1, March 2018, pp. 7~12
ISSN: 2252-8814, DOI: 10.11591/ijaas.v7.i1.pp7-12

Data Mining Techniques for Providing Network Security through Intrusion Detection Systems: a Survey

Prabhu Kavin B, Ganapathy S
School of Computing Science and Engineering, VIT University-Chennai Campus, Chennai, India

Article Info

Article history:

Received Jun 1, 2017
Revised Jan 5, 2018
Accepted Feb 11, 2018

Keywords:

Artificial intelligence
Classification
Clustering
Data mining
Intrusion detection system
Soft computing

ABSTRACT

Intrusion Detection Systems are playing major role in network security in this internet world. Many researchers have been introduced number of intrusion detection systems in the past. Even though, no system was detected all kind of attacks and achieved better detection accuracy. Most of the intrusion detection systems are used data mining techniques such as clustering, outlier detection, classification, classification through learning techniques. Most of the researchers have been applied soft computing techniques for making effective decision over the network dataset for enhancing the detection accuracy in Intrusion Detection System. Few researchers also applied artificial intelligence techniques along with data mining algorithms for making dynamic decision. This paper discusses about the number of intrusion detection systems that are proposed for providing network security. Finally, comparative analysis made between the existing systems and suggested some new ideas for enhancing the performance of the existing systems.

Corresponding Author:

Prabhu Kavin B,
School of Computing Science and Engineering,
VIT University-Chennai Campus,
Chennai, 600 127-India.
Email: lsntl@ccu.edu.tw

1. INTRODUCTION

Now a day's internet has become a part of our life. There are many significant losses in present internet-based information processing systems. So, the importance of the information security has been increased. The one and only basic motto of the information security system is to developed information defensive system to which are secured from an unjustified access, revelation, interference and alteration. Moreover, the risks were related to the confidentiality, probity and availability will have been minimized.

The internet-based attacks were identified and blocked using different systems that have been designed in the past. The Intrusion detection system (IDS) is one of the most important systems among them because they resist external attacks effectively. Moreover, the IDS act as the wall of defense to the computer systems over the attack on the internet. The traditional firewall detects the intrusion on the system but the IDS performance is much better than the firewalls performance. Usually the behavior of the intruders is different from the normal behavior of the legal user, depending upon the behavior the assumption is made and the intrusion detection is done [1]. The computer system files, calls, logs, and the network events are monitored by the IDS to identify the threats on the hosts of the computer. By monitoring the network pockets, the abnormal behavior is detected. The attack pattern is known by finding the possible attack signatures and comparing them. By the known attack signature, the threats are detected easily by the system where as it cannot detect the unknown attacks [2].

An intelligent IDS acts flexible to increase the accuracy rate of the detection. Intelligent IDS are nothing but intelligent computer programs that are located in host or network. By firing the rules of inference

and also by learning the environment, the actions to be performed are computed by the intelligent IDS on that environment [1]. The regular network service is disrupted by transmitting the large amount of data to execute a lower level denial of service attacks. To cause a denial of service attack to the user, the receiver's network connectivity was overwhelmed by creating a specific service request or by sending a large amount of data. The initiation of attack was done by a single sender or the compromised hosts by the attacker and from the latter variant will identify the Distributed Denial of Service (DDoS) [3]. The IDS work same as the Transparent Intrusion Detection System (TIDS) and for the non-distributed attacks the functionality to prevent the attack are provided. The scalability of the traffic processor is achieved by the load balancing algorithm and the system security is achieved by the transparency of nodes. The methodology of anomaly-based attack detection is used in high speed network to detect DDoS attacks, in this method the SDN components are coupled with traffic processor [3].

Among many cyber threats Botnet attack was one the most severe cyber threat. In this attack botmaster is a controlling computer that compromised and remote controlled. Huge numbers of bots were spread over the internet and the botmaster uses the botnet by maintaining under its control. The botnet was used for various purposes by the botmaster, in that few are launching and performing of distributed cyber attacks and computational tasks. The IDS built for botnets are rule based and performance dependant. By examining the network traffic and comparing with known botnet signature the botnet was found in a rule-based botnet IDS. However, keeping these rules updated in the increasing network traffic is more tedious, difficult and time-consuming [4]. Machine-learning (ML) technique is a technique used to automate botnet detection process. From previously known attack signatures a model was built by the learning system. The features like flexibility, adaptability and automated-learning ability of ML is significantly better than the rule-based IDSs. High computational cost is needed for the machine learning based approaches [4].

In this paper, we have discussed about the various types of Intrusion Detection Systems which are used data mining techniques. Rest of this paper is organized as follows: Section 2 provides the related works in this direction. Section 3 shows the comparative analysis. Section 4 suggests new ideas to improve the performance of the existing systems. Section 5 concludes the paper.

2. RELATED WORKS

This section is classified into two major subsections for feature selection and classification techniques which are proposed in this direction in the past.

2.1. Related Works on Feature Selection Methods

Feature selection was the most famous technique for dimensionality reduction. In this the relevant features is of detected and the irrelevant ones are discarded [5]. From the entire dataset the process of selecting a feature subset for further processing was proceeded in feature selection [6]. Feature selection methods are classified into two types, individual evaluation and subset evaluation. According to their degrees of importance feature ranking methods estimate features and allot weights for them. In contrast, build on a some search method subset evaluation methods select candidate feature [4].Feature selection methods is divided into three methods they are wrappers methods, filters methods and embedded methods [5]. An intelligent conditional random field-based feature selection algorithm has been proposed in [7] for effective feature selection. This will be helpful for improving the classification accuracy.

In wrapper method optimization of a predictor is involved as a segment of the selection process, where as in filter method selection the features with self determination of any predictor by relying on the general characteristics of the training data is done. In embedded methods for classification machine learning models was generally used, and then the classifier algorithm builds an optimal subset or ranking features [5]. Wrappers method and embedded method tried to perform better but having the risk of over fitting when the sample size is small and being very time consuming. On the other hand, filter method was more suitable for large datasets and much faster. Comparing with wrappers and embedded methods filters were implemented easily and has better scale up than those methods. Filter can be able to use as a preprocessing step prior trying to other complex feature selection methods. The two metrics of the filter methods in classification problems are correlation and mutual information, along with some other metrics of the filter method like error probability, probabilistic distance, entropy or consistency [5]. In wrapper approach based on specified learning algorithm it selects a feature subset with a higher prediction performance. In embedded as similar as wrapper approach during the learning process of a specified learning algorithm it selects the best feature subset. In the filter approach the feature subset is chosen from the original feature space according to pre-specified evaluation criterions subset using only the dataset. In hybrid approach combining the advantages of the wrapper approach and the filter approach it uses the individualistic criterion and a learning algorithm to rate the candidate feature subsets [8].

In high dimensional applications feature selection is very much important. From the number of original features, the feature selection was the combinatorial problem and found the optimal subset was NP-hard. While facing imbalanced data sets feature selection is very much helpful [9]. Rough set-based approach uses attribute dependency to take away the feature selection, which was important. The dependency measure that was necessary for the calculation of the positive region but while calculating it was an extravagant task [6]. Depend on the particle swarm optimization (PSO) and rough sets, the positive region-based approach has been presented. It is a superintended combined feature selection algorithm and by using the conventional dependency, fitness function was measured for each particle is evaluated. The algorithms figure-out the strength of the selected feature with various consolidations by selecting an attribute with a higher dependency value. If the particle's fitness value is higher than the previous best value within the current swarm (pbest), then the particle value is the current best (gbest). Then its fitness was compared with the population's overall previous best fitness. The article fitness which is better will be at the position of best feature subset. The particle velocities were updated at the last. The dependency of the decision attributes which was on the conditional attributes was calculated by positive region based dependency measure and only because of bottleneck for large datasets it is suitable only for smaller ones [6].

Incremental feature selection algorithm (IFSA) is mainly designed for the purpose of subset feature selection. The starting point is the original feature subset P, in an incremental manner the new dependency function was calculated and required feature subsets are checked. P is the new feature subset if the dependency function P is equal to the feature subset if not it computes a new feature subset. The gradually selected significant features were added to the feature subset. Finally, by removing the redundant features the optimal output is ensured. Then again, the algorithm used the positive region-based dependency measure, and to make it unsuitable for large datasets [6]. Fish Swarm algorithm was started with an initial population (swarm) of fish for searching the food. Here every candidate solution is represented by a fish. The swarm changes their position and communicates with each other in searching of the best local position and the best global positions. When a fish achieved maximum strength, it loses its normal quality after obtaining the Reduct rough set. After all of the fishes have lost it normal quality the next iteration starts. After the similar feature reduct was obtained under three consecutive iterations or the largest iteration condition was reached, then the algorithm halts. Then equivalent rough set-based dependency measure was used in this algorithm and it suffers from the same problem of the large datasets performance degradation [6].

Correlation-based Feature Selection is a multivariate subset filter algorithm. A search algorithm united with an estimation function that was used to evaluate the benefit of feature subsets. The implementation of CFS used the forward best first search as its searching algorithm. Best first search is one of the artificial intelligence search scenario in which backtracking was allowed along with the search path. By making some limited adjustment to the current feature subset it moves through the search space. This algorithm can backtrack to the earlier subset when the explored path looks unexciting and advance the search from there on. Then the search halted, if five successive fully expanded the subsets shows no development over the present best subset [5].

The objective of SRFS is to find the feature subset S with the size d, which contains the representative features, in which both the labeled and unlabeled dataset are exploiting. In this the feature relevance is classified in to three disjoint categories: strongly relevant, weakly relevant and irrelevant features [10-12]. A strong relevant feature was always basic for an optimal or suboptimal feature subset. If the strong relevant feature is evacuated, using the feature subset the classification ability is directly influenced. Except for an optimal or suboptimal feature subset at certain conditions, a weak relevant feature is not always necessary. Irrelevant feature it only enlarges search space and makes the problem more complex, and it doesn't provide any information to improve the prediction accuracy so it is not necessary at any time. Hence all features of strongly relevant and subset features of weakly relevant and no irrelevant features should be included by the optimal feature subset. An in addition supervised feature selection method that uses the bilateral information between feature and class that tend to find the optimal or suboptimal features over fitted to the labeled data, when a small number of labeled data are available. In this case, data mitigation may be able to occur in this problem on using unlabeled data. Therefore, relevance gain considering feature relevance in unlabeled dataset, and propose a new framework for feature selection on removing the irrelevant and redundant features called as Semi-supervised Representatives Feature Selection algorithm is defined. SRFS is a semi-supervised filter feature selection based on the Markov blanket [8]

2.2. Related Works on Classification Algorithms

The combined response composed by the multiple classifiers into a single response was the ensemble classifier. Even though many ensemble techniques exist, for a particular dataset it was hard to found suitable ensemble configuration. Ensemble classifiers are used to maximize the certainty of several classification tasks. Many methods have been proposed, with mean combiner, max combiner, median

combiner, majority voting and weighed majority voting (WMV) whereas the individual classifiers can be connected using any one of these methods [13]. To solve classification and regression problems support vector machines (SVM) is an effective technique. SVM was the implementation of Vapnik's Structural Risk Minimization (SRM) principle which has comparatively low generalization error and does not suffer much from over fitting to the training dataset. When a model performs poor and not located in the training set then it was said to be over fit and has high generalization error [13]. Recently a significant attention was attracted by the multi-label classification, which was motivated by more number of applications. Example include text categorization, image classification, video classification, music categorization, gene and protein function prediction, medical diagnosis, chemical analysis, social network mining and direct marketing and many more examples found. To improve the classification performance by the utilization of label dependencies was the key problem in multi-label learning and how it is motivated by which number of multi-label algorithm that have been proposed in recent years (for extensive comparison of several methods). The progress in the MLC in recent time was summarized. Feature space Dimensionality reduction, i.e. reducing the dimensionality of the vector x is one of the trending challenges in MLC. The dimensionality of feature space can be very large and this issue in practical applications is very important [14]. Many intelligent intrusion detection systems have been discussed in [1] and also briefly described the usage of artificial intelligence and soft computing techniques for providing network security. Moreover, a new intelligent agent based Multiclass Support Vector Machine algorithm which is the combination of intelligent agent, decision tree and clustering is also proposed and implemented. They proved their system was better when compared with other existing systems. Recently, temporal features are also incorporated with fuzzy logic for making decision dynamically [15]. They achieved better classification accuracy over the real time data sets.

2.3. Related works on Clustering and Outlier Detection

Clustering techniques are very useful for enhancing the classification accuracy. Many clustering algorithms have been used in various intrusion detection systems in the past for achieving better performance. Clustering techniques are useful in both datasets such as network trace data and bench mark dataset for making effective grouping [16], [17]. Outlier detection is also useful for identifying the unrelated users in a network. This outlier detection technique is used for identifying the outliers in a network. It can be applied in real network scenario and both datasets such as network trace dataset and the benchmark dataset. Moreover, soft computing techniques are used in these two approaches for making final decisions over the datasets. The existing works [18], [19] achieved better detection accuracy.

3. COMPARATIVE ANALYSIS

Most of the Intrusion Detection Systems have been used data mining techniques such as Clustering, Outlier detection, Classification and data preprocessing. Here, data preprocessing techniques are used to enhance the classification accuracy. Feature selection methods are used to reduce the classification time. This paper describes various types of feature selection which are proposed in this direction in the past. The average performance of the existing classification algorithms is 94% and it has improved into 96% when applied data preprocessing. In addition, the average detection accuracy is reached to 99% when used clustering or outlier detection techniques. Table 1 shows the performance comparative analysis.

Table 1. Comparative Analysis

No.	Author name	Method	Overall Accuracy (%)
1	Srinivas Mukkamala et al [20]	SVM	99.63%
2	Ganapathy et al [21]	IAEMSVM	91.13%
3	Ganapathy et al [1]	IREMSVM	91.26%
4	Soo-YeonJi et al [2]	MLIDS	96%
5	Omar Y. Al-Jarrah et al [4]	RDPLM	99.98%
6	Abdulla Amin Aburomman et al [5]	KNN	91.68%
7	Abdulla Amin Aburomman et al [5]	Ensemble	92.74%
8	VinodkumarDehariya et al [16]	FKM	83.16%
9	UjjwalMaulik et al [16]	GA-FKM	88.46%
10	ChenjieGu et al [16]	IGA-FKKM	93.01%
11	Ganapathy et al [11]	IGA-NWFCM	94.86%
12	J. Ross Quinlan et al [12]	ID3	95.58%
13	Ernst Kretschmann et al [16]	C4.5	96.19%
14	GuoliJi et al [18]	MSVM	98.38%
15	Ganapathy et al [15]	EMSVM	99.10%
16	Ganapathy et al [19]	WDBOD	99.52%

From Table 1, it can be seen that the performance of the method RDPLM perform well than the existing methods and the existing classifier SVM achieved very less detection accuracy than others. This is due to the use of various combinations of methods and the use of intelligent agents.

Figure 1 demonstrates the performance analysis in graph between the top five methods which are proposed in the past by various researchers. Here, we have considered the same set of records for conducting experiments for finding the classification accuracy. Classification accuracy of various methods is considered for comparative analysis.

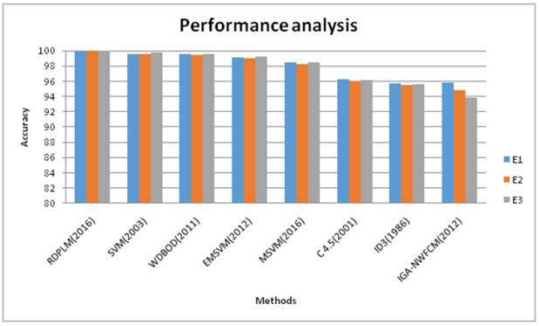

Figure 1. Performance analysis

From figure 1, it can be observed that the performance of the method RDPLM is performed well when it is compared with existing methods. Moreover, the IGA-NWFCM method achieves very less detection accuracy than the other existing algorithms which are considered for comparative analysis

4. SUGGESTION PROPOSED

The performance of the existing systems can be improved by the introduction of intelligent agents and soft computing techniques like fuzzy logic, neural network and genetic algorithms for effective decision over the dataset. In this fast world, time and space are also very important to take effective decision. Finally, can introduce a new system which contains new intelligent agents, neural network for training, effective spatio-fuzzy temporal based data preprocessing method and fuzzy temporal rules can be used for making effective decision and also can detect attackers effectively. This combination is able to provide better performance.

5. CONCLUSION

An effective survey made in the direction of data mining technique-based intrusion detection systems. Many feature selection methods have been discussed in this paper and their importance are highlighted. Classification, Clustering and outlier detection techniques are explained in this paper and also explained how much it is helpful for enhancing the performance. Finally, suggestion also proposed in this paper based on the comparative analysis of the existing systems.

REFERENCES

[1]. S. Ganapathy, K. Kulothungan, S. Muthurajkumar,M. Vijayalakshmi, P.Yogesh, A.Kannan, "Intelligent feature selection and classification techniques for intrusion detection in networks : a survey", *EURASIP Wireless Journal of Communications and Networking*, vol. 2013, pp. 1–16, 2013.

[2]. S. Y. Ji, B. K. Jeong, S. Choi, and D. H. Jeong, "A multi-level intrusion detection method for abnormal network behaviors," *J. Netw. Comput. Appl.*, vol. 62, pp. 9–17, 2016.

[3]. O. Joldzic, Z. Djuric, and P. Vuletic, "A transparent and scalable anomaly-based DoS detection method," *Comput. Networks*, vol. 104, pp. 27–42, 2016.

[4]. O. Y. Al-Jarrah, O. Alhussein, P. D. Yoo, S. Muhaidat, K. Taha, and K. Kim, "Data Randomization and Cluster-Based Partitioning for Botnet Intrusion Detection," *IEEE Trans. Cybern.*, vol. 46, no. 8, pp. 1796–1806, 2016.

[5]. A. A. Aburomman and M. Bin Ibne Reaz, "A novel SVM-kNN-PSO ensemble method for intrusion detection system," *Appl. Soft Comput. J.*, vol. 38, pp. 360–372, 2016.

[6]. P. Teisseyre, "Neurocomputing Feature ranking for multi-label classi fi cation using Markov networks," vol. 205, pp. 439–454, 2016.

[7]. S Ganapathy, P Vijayakumar, P Yogesh, A Kannan, "An Intelligent CRF Based Feature Selection for Effective Intrusion Detection", *International Arab Journal of Information Technology*, vol. 16, no. 2, 2016.
[8]. V. Bolón-Canedo n, I. Porto-Díaz, N. Sánchez-Maroño, A. Alonso-Betanzos, "A framework for cost-based feature selection," *Pattern Recognition, Elsevier*, vol. 47,pp. 2481–726, 2014.
[9]. M. S. Raza and U. Qamar, "An incremental dependency calculation technique for feature selection using rough sets," *Inf. Sci. (Ny).*, vol. 343–344, pp. 41–65, 2016.
[10]. L. Yu, H. Liu, "Efficient feature selection via analysis of relevance and redundancy", *The Journal of Machine Learning Research,* vol.5, pp. 1205–1224, 2004.
[11]. G. H. John, R. Kohavi, K. Pfleger, et al., "Irrelevant features and the sub-set selection problem", *in: Machine Learning: Proceedings of the Eleventh International Conference,* pp. 121–129, 1994.
[12]. B. Grechuk, A. Molyboha, M. Zabarankin, "Maximum entropy principle with general deviation measures", *Mathematics of Operations Research*, vol.34, no. 2, pp. 445–467, 2009.
[13]. Q. Li, Z. Sun, Z. Lin, and R. He, "Author ' s Accepted Manuscript Transformation Invariant Subspace Clustering Reference : To appear in : Pattern Recognition," 2016.
[14]. S. Maldonado, R. Weber, and F. Famili, "Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines," *Inf. Sci. (Ny).*, vol. 286, pp. 228–246, 2014.
[15]. S Ganapathy, R Sethukkarasi, P Yogesh, P Vijayakumar, A Kannan, "An intelligent temporal pattern classification system using fuzzy temporal rules and particle swarm optimization", *Sadhana*, vol. 39, no. 2, pp. 283-302, 2014.
[16]. S Ganapathy, K Kulothungan, P Yogesh, A Kannan, "A Novel Weighted Fuzzy C–Means Clustering Based on Immune Genetic Algorithm for Intrusion Detection", *Procedia Engineering*, vol. 38, pp. 1750-1757, 2012.
[17]. K Kulothungan, S Ganapathy, S Indra Gandhi, P Yogesh, A Kannan, "Intelligent secured fault tolerant routing in wireless sensor networks using clustering approach", *International Journal of Soft Computing*, vol. 6, no. 5, pp. 210-215, 2011.
[18]. S.Ganapathy, N.Jaisankar, P.Yogesh, A.Kannan, " An Intelligent System for Intrusion Detection using Outlier Detection", *2011 International Conference on Recent Trends in Information Technology (ICRTIT)*, pp. 119-123, 2011.
[19]. N Jaisankar, S Ganapathy, P Yogesh, A Kannan, K Anand, "An intelligent agent based intrusion detection system using fuzzy rough set based outlier detection", *Soft Computing Techniques in Vision Science*, pp. 147-153, 2012.
[20]. A. H. Sung and S. Mukkamala, "Identifying Important Features for Intrusion Detection Using Support Vector Machines and Neural Networks", *Department of Computer Science New Mexico Institute of Mining and Technology*, pp. 3–10, 2003.
[21]. S. Ganapathy, P. Yogesh, and A. Kannan, "Intelligent Agent-Based Intrusion Detection System Using Enhanced Multiclass SVM," vol. 2012, 2012.