A Review of Clustering Techniques Based on Machine learning Approach in Intrusion Detection Systems

Ala' Yaseen Ibrahim Shakhatreh [1], Kamalrulnizam Abu Bakar [2]

[1,2] Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia
81310, Johor Bahru, Malaysia

Abstract

False alarm rate and detection accuracy are still challenging issues that are not completely solved yet in the field of Anomaly based Intrusion Detection System (AIDS). The reasons behind these issues vary according to the algorithm and the dataset used to train the IDS. Consequently, dealing with high dimensional data requires an efficient data reduction technique that considerably reduces the dimensionality without any substantial loss in the important features. However, the excessive reduction of features will lead to model some intrusive patterns similarly as normal ones. Indeed, this will result in misclassifications that will increase false negative rate, which degrades the accuracy of detection. This paper concludes many clustering techniques that were previously proposed to solve the inherent IDS problems. Where, the clustering techniques involved in three general aspects namely: data preprocessing, anomaly detection, and data projection/alarm filtering. Eventually, recommendations for future researches followed by the conclusion are depicted at the end of this paper.

Keywords: *Intrusion Detection System, Clustering Techniques, Unsupervised Learning, Detection Rate, False Alarm Rate, Dataset, LVQ, SOM.*

1. Introduction

The nowadays internet has expanded without limits, as many systems moved online to gain the profit of internet marketing. Consequently, the number of security incidents has exploded. Thus, internet security became a growing concern that led many security research oriented organizations to conduct studies to provide an acceptable level of protection against intruders [1]. Intrusion Detection Systems became a necessary complement to the traditional firewalls to ensure data integrity, confidentiality and availability for data transmitted over the network. Relatively, Intrusion Detection Systems are categorized based on two approaches: misuse and anomaly based detection. Misuse is an efficient way to detect known attacks that have known hard coded signatures stored in the signature list. However, any simple variation from the listed signatures will lead to consider such an attack as a legitimate request leading to a high false negative rate. One fact about misuse approach is its low false positive rate, due to its pattern matching techniques with the signature list. However, misuse approach has failed in detecting unknown and 0-day attacks. Thus, the knowledge base must be updated frequently and manually. On the other hand, anomaly based approach depends on establishing normal profile usage and any violation of that normal profile will be considered as an anomalous request. Accordingly, anomaly approach outperforms misuse approach in terms of detection capability of novel and unknown attacks without any advance knowledge which indicates lower false negative rate. Although anomaly detection approach uses machine learning techniques (supervised & unsupervised) to successfully identify unseen attacks, but these techniques tend to generate high false positive rate due to the high dimensionality of datasets used in the training process. Despite anomaly detection approach has low false negative rate, but it is still prone to have false negative alarms because some attacks can be conducted in more than one way, which raises the issue of modeling and the generality of attacks that are not completely solved yet.

In this study our categorization criteria for the connectionist models used in IDSs is based on the functionality of algorithms used. The first group discusses the algorithms used for data preprocessing, the second one discusses the algorithms used for detection process, and the last one is for algorithms used for data projection and alarm filtering. The rest of this paper is organized as follows: section 2.0 shows critical IDSs issues, section 3.0 discusses the variety of clustering algorithms used in IDSs, section 4.0 explores the future trend of IDS, and section 5.0 depicts the conclusion of this paper.

2. Intrusion Detection Systems Issues

All previously proposed solutions mentioned in section 3.0 have addressed the issues or part of the issues that are addressed in this section, attempting to overcome specific limitations in the field of Intrusion Detection System. In general, anomaly based techniques rely on two assumptions: First, the number of normal patterns outweighs the number of intrusive ones; second, the intrusive patterns are qualitatively different from normal patterns [2]. As a result, when using any clustering algorithm to cluster a dataset, some normal patterns mistakenly fall into an anomalous cluster which generates false positive alarms [3].

The problem of increased false negative alarms is that some attack patterns are incorrectly mapped to a normal cluster which contains normal patterns. On the other hand, the problem of false positive alarms is originated from the fact of some normal patterns are incorrectly mapped to a cluster containing anomalous patterns and labeled as anomalous cluster. Furthermore, a key factor in this problem depends on how unsupervised learning techniques adapt with data inputs through its topology, and how they conduct the training process to shape the final topology. Moreover, preprocessing data inputs before training has strong impact on the classification results, where omitting any important feature may affect the detection capabilities and leads to high false alarm rate [3].

On the other hand, an intrusive pattern may fall into a legitimate cluster which leads to a false negative alarm. The reasons behind this misclassification can be the nature of architecture/topology of the clustering algorithm itself and the values of its parameters provided. Additionally, taking into account the first assumption, if the dataset contains considerable amounts of attacking patterns, the clustering technique will consider the cluster in which anomalous patterns fall in as a legitimate cluster. In other words, the IDS can be trained to accept attacks as legitimates [4]. Moreover, polymorphic attacks are still a challenging obstacle to be solved by anomaly approach, because, many attacks are modeled in similar way to normal patterns, indeed this requires dealing with the generality of attacks.

Consequently, this concern will fire up the trade-off between the detection rate and false alarm rate [5]. Furthermore, attacks could not be distinguished from normal patterns because most of critical features in the packet headers which may lead to identify attacks are not utilized.

3.0 Cluster based IDSs

Clustering techniques are the most appropriate choice for dimensionality reduction purposes, especially when a huge multivariate dataset is involved in the training process. For instance, let's take the well-known Self-Organizing Map (SOM) proposed by [6] as a data dimensionality reduction technique that was used either for anomaly detection or data preprocessing in the subsequent literatures in this paper. The high capability of SOM in transforming high dimensional input space onto a very low dimensional neuron space (one or two topological maps) made it preferred for dimensionality reduction purposes. Yet another important feature in SOM and some other clustering techniques is that the ability of identifying outliers presented in the dataset during training phase [7]. Relatively, [8] described SOM as the best choice for dimensionality reduction due to its hard-competitive training approach in the clustering process, which robustly allows dividing the data input to certain number of classes. More detailed information about SOM can be found in [9], [7], and [10].

In general, there are three main aspects in which clustering algorithms can cope with as follows: First, data preprocessing. Second, anomaly detection. Third, data projection and alarms filtering. As our categorization criteria will be based on these three aspects.

3.1 Clustering Algorithms for Data Preprocessing

In [11], Labib and Vemuri assumed that that real-time processing and performance can be achieved using clustering technique specifically by SOM algorithm, by implementing simpler design.

3.1.1 Real-time Data Preprocessing Using Self-Organizing Map

According to the above mentioned assumption in [11], SOM was used in [12] to support online preprocessing for data preprocessing stage. The Distribution Gravity Center (DGC) was used for normalization in the preprocessing stage due to its improvements of the firing behavior of SOM instead of the Euclidean normalization method. Furthermore, the source and destination addresses in the packet headers captured from different protocols were translated into octet vectors to reduce the redundancy in the preprocessed patterns, ignoring the upper two redundant octets in all vectors. Thus, the lowest two octets along with the encoded protocol are involved in the preprocessing process. Moreover, for better pattern clustering, Kohonen Random initialization function was selected over others. The results show better detection capability and the ability to preprocess data in real-time comparing with the original SOM.

3.1.2 IDS based on Self-Organizing Map and K-mean algorithms

S-K algorithm was proposed in [7], to provide better detection and lower false positive arte by combining SOM and K-mean algorithms. Basically, SOM preprocessed the data to produce a number clusters with centers. On the other hand, K-mean algorithm is applied to refine the final results of SOM topology by eliminating grays clusters and remaining black and white ones. The results show that the new algorithm has achieved 92% detection rate and 35% false rate. However, these results were obtained based on testing the algorithm using up to three types of attacks only.

3.1.3 Feature Reduction Using Principal Component Analysis (PCA) Algorithm

The current IDS proposals emphasize on reducing the number of features as inputs to the neural model in order to maintain the performance and the efficiency. However, omitting some feature could result in modeling such an attack similarly to normal pattern, which consequently increases false alarm rate and reduces detection rate. Principal Component Analysis (PCA) algorithm was used as a compacting technique in [13], after all significant features were analyzed without any substantial loss in the information. Therefore, PCA is responsible for keeping the necessary amount of sufficient data for classification and maintaining the performance of the classifier by keeping the number necessary data at its minimum without affecting the effectiveness. As a result, the input vector for every TCP/IP packet to the neural model contains 20 inputs (extracted from 419 inputs of the original TCP/IP inputs) which made it 438 times faster than the original one.

As mentioned before, involving all features can guarantee better and accurate detection, but with more shortcomings on the real-time efficiency. Therefore, in [14] the Principal Component Analysis algorithm (PCA) was used as a feature selection algorithm in order to increase the detection rate along with the accuracy and to decrease the total complexity by reducing the dimensionality of the sample inputs. Moreover SOM was used for clustering and anomaly detection. For the sake of better evaluation, five new attack types were added to the typical attacks presented in the obtained 10% of KDD Cup 1999 as a sample dataset. The results of the proposed MPCA-MSOM algorithm shows better detection rate and lower false positive rate, 97.0% and 2.2% respectively comparing with the original SOM, K-NN, and RoughSet algorithms. As a result, the MPCA-MSOM algorithm not only proved to have better detection rate and false positive rate, but provided better attack classification.

3.1.4 Prior Knowledge for Better Traffic Characterization Using SOM and Learning Vector Quantization (LVQ)

In [5], eleven SOMs, one for every attribute selected to validate the variation of each attribute of the packet headers separately over time window, aiming to identify the main features of every attribute. While the second layer consisting of 6 by 6 SOM and LVQ assigned to each SOM in the first layer to correlate the first layer information between the attributes and classifies them among three classes: Normal, Attack or Indefinite. The last layer decides whether the vectors in the indefinite class are Normal or Attack using SOM. Furthermore, the LVQ network which is assigned to each SOM in the second level is to make the final classification of Indefinite to either Normal or Attack. The results show that the detection rate decreased by 19% which rates below comparing with: clustering, K-NN, SVM, and SOM Hierarchy, because user-to-root attack is difficult to model using the characteristics of TCP/IP traffic only.

3.1.5 Detecting 0-Day Attacks Using CMLHL Connectionist Model

Utilizing CMLHL connectionist model among multi-agent system in [15] to enhance pattern classification, for the sake of detecting 0-day attacks and other attacks conducted through SNMP protocol. Additionally, the architecture consists of one central IDS agent for anomaly detection using the connectionist model served by several sniffer agents distributed among all network segments sniffing and preprocessing packet headers captured. On important aspect of this approach is that it shows the temporal relationship between packets within time dimension. However, this architecture creates a bottleneck situation due to the huge payloads flocking from sniffer agents to the IDS agent, which requires the IDS agent to be on a huge calculus power machine. Moreover, the scope of attacks is related to SNMP attacks only.

3.1.6 Modeling Real-World Traffic to Generate Syntactic Dataset

A framework was developed in [16], to generate synthetic traffic based on HTTP protocol to measure the improvements of the generated synthetic traffic over the 10% of KDD Cup 1999 (containing HTTP connections only) and compare the results with the results of real-world traffic when testing them using K-mean and SOM based IDSs. Moreover, the traffic was generated in tcpdump format, and BRO network analyzer was customized to extract 41 features. However, only 6 features were involved in this testing namely: duration of the connection; protocol; service; connection status; total bytes sent to

destination host; total bytes sent to source host. After testing the three datasets using the K-mean clustering algorithm, it can be concluded that the generated synthetic dataset has more similarities to the real-world traffic (generated from Faculty of Computer Science server Locutus in Dalhousie University) than KDD Cup 1999.

On the other hand, when testing the datasets using SOM based IDS developed by [15], by applying two hierarchy levels of SOM. At the first level each feature of the six features is assigned to a SOM to be trained aiming to encode temporal relationships among the features. The second level, the information from the first level is combined to be represented and labeled. The results concluded that the SOM trained on the synthetic dataset shows improvements and more similarities to the SOM trained on the real-world dataset than the one that trained on KDD99.

3.2 Clustering Algorithms for Anomaly Detection

Supervised and unsupervised clustering algorithms were utilized and improved to enhance detection capabilities and false alarm rate.

3.2.1 Combining SOM Algorithms for Scalable IDS

In order to avoid the high complexity of the training process of the original SOM, the most proper SOM type is selected and assigned to network node in [18], that in order to cope with real-time requirements by reducing the training cost. The selection of SOM type is based on the following criteria: scalable SOM is used if the number of zero-elements is greater than 40%, and there is no memory limitation. Scalable SOM with compressed vectors is used if the memory limitation matters. If the number of features is lower than 10, the original SOM is used. Otherwise, if it is very high more than 500 or the memory resources are very limited or the visualization is not needed, then HSOM is used because it is fast. Additionally, if there is no memory limitation and the CPU has limited resources, then GHSOM is used. Moreover, if the importance of node is very high, then the original SOM is involved. Furthermore, fast winner search technique is used whenever real-time processing is important. As a result, the optimal training cost can be achieved based on these criteria.

3.2.2 Multi-agent IDS based on Clustering Algorithm

Yet another approach to enhance detection capabilities through multi-agent system or distributed IDS as applied in [15], [19], and [20]. Thus, enhancing pattern classification in [15] through CMLHL connectionist model is described in section 3.1.5. While in [19], the IDS is composed of distributed agents (one for each network node) administrated by administrator agent which uses clustering technique based on new growing Self-Organizing Map model to provide flexibility and modularity in detecting anomalies. Moreover, the distributed agents detect the anomalies based on the shared knowledge in the administrator agent. After training and testing the system using wide scope of attacks, about 22 attack types in addition to normal patterns presented in KDD Cup 1999, the results showed 90.79% detection rate which is lower than other works due to wider scope of attacks and features involved.
Combining Case-Based Reasoning (CBR) and neural network through multi-agent system in [20], to increase the detection and projection capabilities of SNMP related anomalies. The architecture consists of six evolving agents that can dynamically learn and adapt using the neural model. The analyzer agent is a CBR-BDI agent that applies neural model within its adaptation stage for analyzing the preprocessed segments to allow the projection of network traffic. Furthermore, several neural models namely: PCA, CCA, MLHL, and CMLHL were applied on a dataset obtained from SNMP traffic using 5 features for comparison reasons. As a result, CMLHL model found to have the best projection and attack identification comparing with MLHL, PCA, and CCA models. Moreover, with respect to CCA, the best results were depicted based on Standardized Euclidean Distance.

3.2.3 IDS based on Dynamic Self-Organizing Map DSOM

Anomaly clusters can be identified by the cluster of normal ratio. Thus, in [21], [22], and [2], improving pattern classification is through the use of dynamic SOM (DSOM) with other clustering algorithms to make the detection independent of the centers of clusters. Apart from traditional clustering by which the accuracy of detection is degraded by the use of simple distance metric, the growing DSOM and Ant Colony Optimization algorithm (ACO) can control clustering efficiency in [21]. Additionally, in the detection stage, Posteriori probabilities make it much independent to increase the efficiency in detecting unknown attacks. Then, after initializing ants, each ant selects an object randomly, and picks it up or moves or drops it based on the probability of each action in DSOM. In the second stage, clusters are gathered from DSOM's output preparing for labeling the obtained clusters. As the detection stage is based on Bayes theorem due to its low fault rate. Eventually, using KDD dataset for testing, the experimental results show better detection rate and false positive rate comparing with K-NN and SVM based IDS.

On the other hand, the same DSOM was used in [22] followed by Swarm Intelligence (SI) clustering for the

same purpose in [21], but in hope to more accurate detection classifier to enhances the detection rate and false positive rate. The results of this approach show better performance in detecting intrusions comparing with LGP, SVM, KNN, and DT. As a result, from these two works, we can conclude that Swarm Intelligence cope better with Dynamic SOM than ACO in quality of clustering and intrusion detection accuracy.

Additionally, after obtaining clusters by Dynamic SOM (DSOM) in [2], Bayesian classification algorithm is optimized as the detection algorithm, which has the least fault rate among other classification algorithms. Furthermore, Bayes theorem makes the detection process independent of the center of clusters, which increase the accuracy of detection. Using dataset D obtained from KDD Cup 1999 which consists of 1% to 1.5% intrusions, and 98.5% to 99% normal patterns for testing and evaluating the proposed IDS. Consequently, the results show higher detection rate and lower false positive rate comparing with Cluster, K-NN, and SVM based IDSs.

3.2.4 Hierarchical Clustering Based IDS

Hierarchical cluster techniques were introduced in [10], [23], [24], [25], and [26] to address the main limitations of traditional clustering techniques. Where, the static architecture is imposed by establishing a fixed topology in advance before the training stage. Moreover, the topology maintains many empty and unnecessary clusters which degrade the efficiency by increasing the complexity of the IDS. Thus, the input vectors are not faithfully represented in the traditional clustering topology, which results in lower detection accuracy and more false alarm rate.

However, in [10], Growing Hierarchical SOM (GHSOM) is optimized for better classification. Considerably, GHSOM consists of several SOMs arranged in layers growing during the training process along with the number of layers and neurons of maps to automatically adapt with data inputs can faithfully represent input vectors. Moreover, a metric based on entropy for symbolic values together with numerical values (by Euclidean distance) is presented in GHSOM in order to involve three pivotal symbolic features namely: protocol type, service, and flag of status in the training stage. After training the model using KDD Cup 1999 dataset, the results of this GHSOM were compared with SOM and K-Map algorithms. The experiments showed detection rate: 99.98, 81.85, and 99.63% respectively and in terms of false positive rate: 3.03, 0.03, and 0.34% respectively, which indicates better detection capabilities with an acceptable FPR.

The growing hierarchical self-organization graph (GHSOG) is proposed in [23], to overcome the limitations of static topology and the lack of representation of hierarchical relations between data inputs in SOM. On the other hand, GHSOM has also some limitations related to the static topology, where each map is initiated with 2x2 neurons, forcing 2-D rectangular grid of map adding many neurons without necessity, taking the map far from the optimal number of neurons. Consequently, GHSOG is based on establishing map topology according to data inputs faithfully, by reflecting data inputs as faithfully as possible. KDD Cup99 was used for training and evaluating the model. Moreover, to avoid preprocessing traffic data in consecutive quantitative way, each qualitative feature is replaced with binary vector consisting of many binary features, allowing using the Euclidean distance function. By this, the number of features increased from 41 to 118. The results show lower detection rate about 90.68% and more FPR comparing with K-Map, SOM, and SOM (DoS). However, the mentioned works used only 3 attack types from KDD Cup99, while in this work 38 attack types were used, where 15 are new and unknown attacks. Furthermore, this model can cope with real-time IDS due to its lower complexity by utilizing lower number of neurons than other works.

Enhanced SVM is used to provide better performance and generalization accuracy in [24]. In this work hierarchical clustering analysis is used through Dynamic Growing Self-Organizing Tree (DGSOT) to cluster huge dataset efficiently and to overcome the limitation of time consuming training phase of SVM. Considerably, DGSOT helps in SVM training by finding the best qualified boundary points between two classes. This work has contributed in: First, reducing SVM training time. Second, the enhanced SVM is proved to be faster than the original. Third, in terms of false positive FP and false negative FN rates this approach outperforms random selection and Rocchio Bundling on a benchmark dataset. After using DARPA dataset, involving DOS, U2R, R2L, and Probe attacking patterns, the results show enhancements over Random Selection, pure SVM, and SVM + Rocchio Bundling in terms of detection rate and FPR. However, this approach still shows low accuracy for detecting U2R and R2L attacks 23% and 43% respectively, which indicates high false alarm rates for these attacks.

The evolution of clustering techniques is still at its peak, leading to innovate new elegant techniques that take the advantages of more than one clustering algorithm. This approach can be seen in this paper through the evolution of Self-Organizing Map algorithm (SOM). The new upgraded SOM models led to better detection rate and lower false alarms. Moreover, real-time efficiency was taken into account in growing hierarchical related models as applied in [10], [23], and [24]. On the other hand, Growing Hierarchical Recurrent SOM (GH-RSOM) was used for efficient clustering in phoneme recognition in [25]. The contribution in this work is to make a hierarchical model that composed of independent RSOMs; each one is allowed to grow during the unsupervised learning process until the quality of data representation is met. Moreover,

the best matching unit is selected through the difference vector defined by each map. And the adaptation of the weight of the map is defined by the difference vector as well. Eventually, comparing the proposed model with GHSOM model, the proposed algorithm (GH_RSOM) provides better classification rates of phoneme recognition. From our perspective, this model is applicable for intrusion detection system to provide better attack classification.

A new Probabilistic Self-Organizing Graph (PSOG) algorithm is proposed in [26], to provide better classification capabilities. And the topology of Self-organizing neural model is adapted to reflect the internal structure of inputs distribution rather than being fixed during the learning phase as applied in other works. Moreover, each unit of the resulted self-organizing graph is a mixture component of Gaussians (MoG). Furthermore, the corresponding update equations are obtained from stochastic approximation framework, as it is used to learn both mixture and the topology. As each probabilistic mixture of multivariate Gaussian components is associated with a neuron or unit. For evaluation purposes, four uniform distributions were selected to show graphically how PSOG model can adapt its topology according to the structure of input distributions. From the results, PSOG shows better adaptation to its inputs and can perform classification better than other static models, because other static models do not learn their topologies.

3.2.5 Novelty Detection Using Clustering Techniques

In this section, two studies were conducted to achieve novelty detection based on combining SOM with Genetic algorithm to produce GSOMS in [27], and comparing SOM-L with One-class Learning Vector Quantization (OneLVQ) in [28] in terms of classifying novel patterns during training.

A combination of SOM and Genetic algorithms is to form Genetic SOM clustering algorithm order to increase the effectiveness of training process in [27]. Moreover, the role of genetic algorithm is for training the synaptic weights of SOMs. In other words, the adjustment of the SOM synaptic weights is conducted through GA instead of traditional learning rules. Specifically, a chromosome of the GA represents possible combinations of synaptic weights of SOM neurons. In previous studies, detecting novel attacks was based on knowledge learnt from labeled data. Furthermore, detecting novel attacks requires training with new labeled data samples. However, this solution (training with new labeled data) is very expensive when dealing with huge network data. In order to evaluate the proposed mode, KDD Cup99 was used, and the 41 features of each connection were divided into 4 categories according to their data types as follows: Strings, Boolean, Count (integer), and Rate (float) types. With the optimal K representative value which equals to 40, GSOMC model achieved 80% detection rate and 1.9% false alarm rate. The low detection rate explains that some attacks presented in the KDD Cup 1999 dataset or in the real-world internet are stealthy and difficult to model using Euclidean distance measure used in many techniques.

In [28], two methods were proposed to detect novel patterns. First, presenting a scheme of SOM-L boundary to determine local threshold. Second, modifying the learning rule of one-class Learning Vector Quantization (OneLVQ) to allow one to keep codebook vectors far from novel patterns as much as possible. A key factor in this work is utilizing the novel patterns presented in the training dataset. According to [29], novel patterns in the training can be utilized to achieve high classification performance. Furthermore, novelty detection means proper generalization to characterize patterns from normal class. In the meanwhile, specialization means excluding patterns from all other classes [30]. Consequently, proper balancing between these two concepts can achieve classification performance. Unlike the original LVQ, OneLVQ is based on LVQ learning rule, but it assigns the codebooks to one class only (the normal class) rather than many classes as in the original LVQ. Accordingly, the error rate is minimized in which the codebooks are forced to be located near the normal patterns and far from novel ones. After testing the two models, OneLVQ correctly classified all three regions O1, O2, and O3 without any misclassification, outperforming the SOM-L method which failed in recognizing large part of normal patterns. As a conclusion, OneLVQ outperform SOM-L in terms of novelty detection and utilizing novel patterns during training phase.

3.3 Clustering Algorithms for Data Projection and Alarm Filtering

Traditional IDSs tend to generate huge amount of alarms during the detection stage, which exhausts system administrator by rendering a large amount of unserious alarms. Fittingly, these enormous amounts of alarms are filtered and utilized to reduce false positive rate and to increase detection accuracy in [3], [31], and [32]. A data mining technique based on Growing Hierarchical Self-Organizing Map (GHSOM) which adjusts its topology during the learning process according to the inputs data (alarms) to reduce false positive alarms and to assist system administrators in analyzing alarms generated by the IDS. The proposed algorithm aims to explore the hidden structure of alarm data, and to uncover false alarms (FP & FN) hiding in normal clusters. Considerably, a data sample consisting of 1849 data patterns including 6 web attack scenarios were tested using the proposed approach. The results show that the proposed algorithm outperforms SOM algorithm in terms of both false positives and false

negatives which reduced from 15% to 4.7% and from 16% to 4% respectively.

A huge amount of alerts generated from the IDS are correlated to such categories to remove the unnecessary or unserious alerts and to make them readable for the administrators in [31]. A major limitation in the previous solutions of alert correlation is that the methods used led to increase false positives. In order to cope with this problem, in this study selected features from alerts are fed to the SOM to provide better correlation. The selected features should be able to identify the behavior of the traffic. Source/destination IP, target port, source IP of the current alert, target IP of the previous alert, sensor ID, and signature ID are capable to identify the intention of the traffic. Accordingly, if two alerts have the same six features, they will be clustered in the same neuron in SOM. Relatively, if two alerts have five similar features, they will be clustered in the vicinity of the six features neuron, and so on until two features are matched. This model can certainly help system analysts in identifying intrusions by concentrating on the groups of alerts that are relevant with each other.

A detection and prevention layer from SQL-Injection attacks through an anomaly visualization agent was proposed in [32], as a complement to an existing IDS called SCMAS. Furthermore, SCMAS has been upgraded by adding new agent (visualizer), its main functionality is to compliment the classification of this attack by improving the performance of classification. Three types of projection models were applied in the anomaly agent namely, Principal Component Analysis (PCA), Curvilinear Component Analysis, and Cooperative Maximum Likelihood Hebbian Learning (CMLHL) as their results were compared with each other. In order to test the model, SQLMAP 0.5 was used to generate malicious queries using all possible types of SQL-injection attacks. Totally, a sample of 1000 normal and malicious patterns was selected as a dataset. As a result, CCA was proved to have the best classification by grouping normal patterns separately from anomalous ones. On the contrary, PCA and CMLHL mix normal patterns with anomalous ones which increase false alarms.

4. Recommendations for Future Researches

At the early time of intrusion detection system, misuse approach was applicable due to the limited number of attacks at that time. For instance, SNORT IDS was widely and successfully applied to provide sufficient protection. However, the exponential growth of new attacks has pushed the researchers to move forward to anomaly approach, due to the insufficiency of misuse approach in detecting unknown attacks. Moreover, based on the comprehensive survey in [33], Soft Computing methods (SC) consisting of Fuzzy Logic (FL), Artificial Neural Network (ANN), Probabilistic Reasoning (PR), and Evolutionary Computing played a significant role in improving the accuracy of detection and false alarm rate through three approaches: consecutive combinations, ensemble combinations, and hybrid combinations. Furthermore, ensemble combinations of several SC methods in parallel as described in figure 1, was proved to be more robust than consecutive and hybrid combinations.

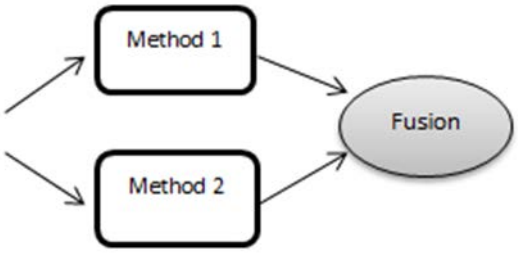

Fig. 1 an ensemble strategy.

Our recommendations for future research are represented in optimizing SC methods in ensemble approach among five stages namely: data preprocessing, features reduction, clustering, training, and classification. However, our main objective in our research is to provide applicable solution in the real-world internet by increasing the scope of attacks along with detecting novel attacks in the detection process with more accuracy. Furthermore, hierarchical clustering techniques were proved to be more robust than other techniques in terms of clustering accuracy and complexity when the number of attacks involved is higher. Relatively, fuzzy classification with rule generation based on partition of overlapping areas was proved to be the most accurate in attacks classification in [33] about 100% accuracy. Consequently, clustering and classification outputs can be more accurate if all important features of connections are involved. However, in order to avoid the high computational cost of the whole process, an efficient feature reduction technique can be optimized without any substantial loss in the important features that may lead to increase the accuracy of detection.

5. Conclusion

In this paper, we walked through the development of anomaly based intrusion detection systems during the recent years. As several supervised and unsupervised clustering techniques were optimized resulting in more elegant techniques that provided more detection accuracy and lower false alarm rate. Moreover, the newly proposed techniques tend to avoid the creation of unnecessary neurons in the training process to faithfully represent data inputs as applied in hierarchical clustering. Furthermore,

this restriction in creating neurons significantly contributes in reducing the complexity of the training process and producing more accurate topologies. Since, our main concern in our research is to increase the quality of clustering and attacks classification for larger scope of attacks. Additionally, increasing the identification rate of novel patterns in the training process as well.

References

[1] A. Bashah Mat Ali, A. Yaseen Ibrahim Shakhatreh, M. Syazwan Abdullah, and J. Alostad, "SQL-injection vulnerability scanning tool for automatic creation of SQL-injection attacks," in Procedia Computer Science, 2011, vol. 3, pp. 453-458.

[1] A. Bashah Mat Ali, A. Yaseen Ibrahim Shakhatreh, M. Syazwan Abdullah, and J. Alostad, "SQL-injection vulnerability scanning tool for automatic creation of SQL-injection attacks," in Procedia Computer Science, 2011, vol. 3, pp. 453-458.

[2] Y. Feng, K. Wu, Z. Wu, and Z. Xiong, "Intrusion Detection Based on Dynamic Self-organizing Map Neural Network Clustering 2 Intrusion Detection Based on DSOM Clustering," in Proceedings of ISNN 2005, pringer-Verlag Berlin Heidelberg, 2005, pp. 428-433.

[3] N. Mansour, M. I. Chehab, and A. Faour, "Filtering intrusion detection alarms," Cluster Computing, vol. 13, no. 1, pp. 19-29, Sep. 2009.

[4] G. Giacinto, "Detection of Server-side Web Attacks," in JMLR: Workshop and Conference Proceedings 11, 2010, vol. 11, pp. 160-166.

[5] A. Carrascal, J. Couchet, E. Ferreira, and D. Manrique, "Anomaly Detection using prior knowledge : application to TCP / IP traffic," IFIP International Federation for Information Processing, Artificial Intelligence in Theory and Practice, vol. 217, pp. 139-148, 2006.

[6] T. Kohonen, "uni Out o t cessi s," Biological Cybernetics, vol. 69, pp. 59-69, 1982.

[7] W. Huai-bin, Y. Hong-liang, X. Zhi-jian, and Y. Zheng, "A Clustering Algorithm Use SOM and K-Means in Intrusion Detection," in 2010 International Conference on E-Business and E-Government, 2010, no. 2007, pp. 1281-1284.

[8] V. K. Pachghare, V. a Patole, and D. P. Kulkarni, "Self Organizing Maps to Build Intrusion Detection System," International Journal of Computer Applications, vol. 1, no. 8, pp. 1-4, Feb. 2010.

[9] D. V. Raje, H. J. Purohit, Y. P. Badhe, S. S. Tambe, and B. D. Kulkarni, "Self-organizing maps: a tool to ascertain taxonomic relatedness based on features derived from 16S rDNA sequence.," Journal of biosciences, vol. 35, no. 4, pp. 617-27, Dec. 2010.

[10] E. J. Palomo, E. Domínguez, R. M. Luque, and J. Muñoz, "An Intrusion Detection System Based on Hierarchical," in Proceedings of the International Workshop on Computational Intelligence in Security for Information Systems CISIS'08, 2009, vol. 53/2009, pp. 139-146.

[11] K. Labib and R. Vemuri, "NSOM : A Real-Time Network-Based Intrusion Detection System Using Self-Organizing Maps," Networks and Security, 2002.

[12] M. Angel and P. Del, "Towards an Intelligent Intrusion Detection System based on SOM Architectures," Applied Sciences, pp. 1-13, 2006.

[13] I. Lorenzo-fonseca, F. Maciá-pérez, and F. J. Mora-gimeno, "Intrusion Detection Method Using Neural Networks Based on the Reduction of Characteristics," in IWANN '09 Proceedings of the 10th International Work-Conference on Artificial Neural Networks: Part I: Bio-Inspired Systems: Computational and Ambient Intelligence, 2009, pp. 1296-1303.

[14] J. Bai, Y. Wu, G. Wang, S. X. Yang, and W. Qiu, "A Novel Intrusion Detection Model Based on Multi-layer Self-Organizing Maps and Principal Component Analysis," Proceedings of ISNN 2006, Springer-Verlag Berlin Heidelberg, pp. 255 - 260, 2006.

[15] E. Corchado, Á. Herrero, and J. M. Sáiz, "A FEATURE SELECTION AGENT-BASED IDS," Symposium A Quarterly Journal In Modern Foreign Literatures, 2007.

[16] H. G. Kayac and N. Zincir-heywood, "Analysis of Three Intrusion Detection System Benchmark Datasets Using Machine," in Proceedings of the IEEE ISI 2005, 2005, pp. 362 - 367.

[17] H. G. Kayacik, A. N. Zincir-Heywood, and M. I. Heywood, "On the capability of an SOM based intrusion detection system," in Proceedings of the 2003 IEEE IJCNN, Portland, USA, July 2003, 2003, vol. 3, pp. 1808-1813.

[18] S. Albayrak, C. Scheel, D. Milosevic, and A. Muller, "Combining Self-Organizing Map Algorithms for Robust and Scalable Intrusion Detection," in International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06), 2005, pp. 123-130.

[19] E. J. Palomo, E. Dom, R. M. Luque, and J. Mu, "A Self-Organized Multiagent System for Intrusion Detection," in Agents and Data Mining Interaction, 2009, pp. 84-94.

[20] Á. Herrero and E. Corchado, "Agents and Neural Networks for Intrusion Detection," International Workshop on Computational Intelligence in Security for Information Systems 2008, pp. 155-162, 2009.

[21] Y. Feng, J. Zhong, Z.-yang Xiong, C.-xiao Ye, and K.-gui Wu, "Network Anomaly Detection Based on DSOM and ACO Clustering," in Springer-Verlag Berlin Heidelberg, Part II, LNCS 4492, 2007, pp. 947-955.

[22] Y. Feng, J. Zhong, Z.-yang Xiong, C.-xiao Ye, and K.-gui Wu, "Intrusion Detection Classifier Based on Dynamic SOM and Swarm Intelligence Clustering," Springer Science+Business Media B.V, pp. 969-974, 2008.

[23] E. J. Palomo and D. L, "Hierarchical Graphs for Data Clustering," in IWANN '09 Proceedings of the 10th International Work-Conference on Artificial Neural Networks: Part I: Bio-Inspired Systems: Computational and Ambient Intelligence, 2009, pp. 432-439.

[24] L. Khan, M. Awad, and B. Thuraisingham, "A new intrusion detection system using support vector machines and hierarchical clustering," The VLDB Journal, vol. 16, no. 4, pp. 507-521, Aug. 2006.

[25] C. Jlassi, N. Arous, and N. Ellouze, "The Growing Hierarchical Recurrent Self Organizing," in NOLISP 2009, LNAI 5933, Springer, 2010, pp. 184-190.

[26] E. López-rubio, J. M. Ortiz-de-lazcano-lobato, and M. C. Vargas-gonzález, "Probabilistic Self-Organizing Graphs," in IWANN 2009, Part I, LNCS 5517, 2009, pp. 180-187.
[27] C. C. Lin and M. S. Wang, "Genetic-clustering algorithm for intrusion detection system," International Journal of Information and Computer Security, vol. 2, no. 2, p. 218, 2008.
[28] H.-joo Lee and S. Cho, "SOM-Based Novelty Detection Using Novel Data," in Proceedings of IDEAL 2005, 2005, pp. 359-366.
[29] Y. Zhao and Z. Wang, "[Support vector data description for finding non-coding RNA gene].," Sheng wu yi xue gong cheng xue za zhi = Journal of biomedical engineering = Shengwu yixue gongchengxue zazhi, vol. 27, no. 4, pp. 779-84, Aug. 2010.
[30] D. Fisher, "Machine Learning, Special Issue on Unsupervised Learning, 1?? (to appear)," Computer, vol. 5, pp. 1-29, 2001.
[31] M. Kumar, S. Siddique, and H. Noor, "Feature-based alert correlation in security systems using self organizing maps," in Proceedings of SPIE, 2009, no. Id, pp. 734404-734404-7.
[32] Á. Herrero, C. I. Pinzón, E. Corchado, and J. Bajo, "Unsupervised Visualization of SQL Attacks by Means of the SCMAS Architecture," in Trends in PAAMS, AISC 71, Springer, 2010, pp. 713-720.
[33] C. Langin and S. Rahimi, "Soft computing in intrusion detection: the state of the art," Journal of Ambient Intelligence and Humanized Computing, vol. 1, no. 2, pp. 133-145, Apr. 2010.

Ala' Yaseen Ibrahim Shakhatreh obtained his master degree in Information Technology from Universiti Utara Malaysia (UUM), and currently he is a PHD student in the Department of Computer Systems and Communications of Computer Science and Information Systems Faculty at the Universiti Teknologi Malaysia. His research area is in network security (Intrusion Detection System) and penetration testing. As he is supervised by Assoc. Prof. Kamalrulnizam Abu Bakar.

Kamalrulnizam Abu Bakar obtained his PhD degree from Aston University (Birmingham, UK) in 2004. Currently, he is an Associate Professor in Computer Science at Universiti Teknologi Malaysia (Malaysia) and member of the "Pervasive Computing" research group. He involves in several research projects and is the referee for many scientific journals and conferences. His specialization includes mobile and wireless computing, information security and grid computing.