ANALYSIS OF KNN, C5.0 AND ONE CLASS SVM FOR INTRUSION DETECTION SYSTEM

Meesala Shobha Rani1, Sumathy.S
VIT University, Vellore, Tamilnadu.

Email: meesalashobha.rani2016@vitstudent.ac.in

Received on 25-10-2016 Accepted on 02-11-2016

Abstract:

Cyber security attacks have become increasingly sophisticated and complex. Intrusion detection is one of the major challenge in computer security that aims to detect unseen attacks in order to protect internal networks. Intrusion detection system compact with huge quantity of data which is a crucial task to keep good quality of features and remove the redundant and irrelevant features. An intelligent intrusion detection system combining ensemble classifier for classifying abnormal and normal actions in the computer network is presented. In this work, K-Nearest neighbourhood algorithm is used to determine the best optimal subset. The misuse detection model is built based on the C5.0 algorithm. Further, anomaly detection model is implemented by one-class Support Vector Machine (SVM) to detect the anomalies. Integration of ensemble algorithms helps achieve better classification. Based on the exhaustive analysis carried out, we propose that on incorporating, NSL-KDD Dataset, which is a well known benchmark for intrusion detection dataset, the system will result better in terms of anomaly detection rate, false positive, false negative, true positive, true negative, and f-score.

Keywords: Anomaly detection, C5.0 Algorithm, Intrusion Detection System, K-NN, Misuse detection, one-class SVM.

1. Introduction

A proper intrusion detection system when deployed in an organization can avoid threats and vulnerabilities. Intrusion detection is the art of detecting unsuitable, improper, or abnormal activity both internally and externally. In general intrusion detection methods are characterised as misuse detection and anomaly detection. The misuse detection method detects attacks built on the identified attack signature. It is valid in detecting recognized occurrences with fewer faults and cannot identify recently created occurrences of attacks that do not have related behaviour to the known attacks. In contrast, anomaly detection method detects the unknown attacks. It analyses the present actions...
with the regular profiles and broadcasts important abnormalities as intrusions. Anomaly detection method can be useful for identifying new attack patterns. Dimensionality reduction plays major phase in machine learning as it deals with high dimensional data. The feature selection is one of the part of dimensionality reduction that is also identified as better subset selection and used to reduce and eliminate redundant and unrelated features[11][20].

It is observed that misuse detection does not recognize unknown attacks and anomaly detection does not recognize known attack. Presently, it is required to handle large data sets and ensure that the system handles both known and unknown attacks after classifying the data reducing the dimensionality to improve the detection ratio as key focus. To solve the limitations of the two existing intrusion detection methods, intelligent ensemble intrusion detection (IDS) method which combines KNN, C5.0 and One-Class SVM anomaly detection in order to attain better detection rate and improved classification with low false alarm is discussed.

The paper is structured as follows: Section 2 presents the existing techniques of hybrid intrusion detection system. Section 3 discusses the detailed description of the proposed hybrid intrusion detection approaches and section 4 presents the conclusion.

2. Review of Related work

Extensive work is being carried out for hybrid intrusion detection system and their advantages and limitations are discussed [20].The anomaly detection based on entropy of features and SVM is discussed. The results are implemented and compared individually [2]. The hybrid method achieves high performance compared to the individual methods in terms of accuracy, but it is not dynamic to decide whether it is susceptible to attacks or it causes high false alarms. Hybrid IDS use combination of Packet Header Anomaly Detection (PHAD) and Network Traffic Anomaly Detection (NETAD) for anomaly detection. The hybrid IDS is much powerful than the signature based IDS [4].A multi-start meta-heuristic method and genetic algorithm are used for anomaly detection in huge amount of datasets. The motivation is taken from negative selection based detector generation. This method shows a better accuracy in generating a suitable number of detectors compared to the other machine learning algorithms like NB (Naïve Bayes), J48 (Decision tree), FBNN (Multilayer Feedback Neural Network), Bayes Network (BN), Bayesian Logistic Regression (BLR), Radial Basis Function Network (RBFN)[5].

A hybrid method C5.0 and SVM, achieves better performance compared to the individual SVM and have evaluated using DARPA dataset [6].The integration of clustering, feature selection, SVM for intrusion detection system. This approach provides enhanced performance in terms of accuracy in comparison to the other NIDs. It only detects DOS
and Probe attacks and does not detect U2L and R2L attacks [7]. A Hidden Markov Model (HMM) and C5.0 are combined to achieve better accuracy in comparison to the HMM. The hybrid method reduces the limitations of HMM algorithm [10]. The combination of the misuse detection method uses C4.5 algorithm and anomaly detection method uses one-class SVM. The results are compared with NSL-KDD datasets [11]. A Hidden Naive Bayes (HNB) model for dimensionality reduction removes redundant and irrelevant features. This method achieves better in terms of detection rate, error rate and misclassification cost. This method is compared with Naive Bayes model, extended Naive Bayes model and the KDDCUP1999 winner [12].

An efficient feature removal method for Intrusion detection system using gradual feature removal method, combination of clustering method, ant colony algorithm and SVM is discussed [14]. The ensemble of Support Vector Machine (SVM), Decision Tree (DT) and simulated annealing (SA) for determining feature selection is deployed. This method achieves high accuracy in comparison to the hybrid processes of DT, SA and feature selection. The hybrid particle swarm optimization (PSO), SVM and feature selection, only DT, only SVM are used to simulate the results [15]. A new method for fault detection and diagnosis using 1-class SVM and SVM-recursive feature elimination is presented. This method is based on non-linear distance metric determined in feature space. This method achieves better performances in terms of false alarm rates, detection latency and fault detection rates in comparison to the conventional techniques such as PCA and DPCA [16]. The integration of K-Means, C4.5 methods are used for anomaly detection. This method achieves better performance in comparison to the K-Means, ID3, Naïve Bayes, K-NN, and SVM [18].

An integration of data filtering, supervised or unsupervised method for identifying network attacks with an approach to provide better detection rate and low false alarms is presented [19].

An integration of NBC, NB for attack detection method gives good recall and precision, but lacks in the detection performance. Single algorithm is not effective in detecting minority attacks, and hence combining of multiple algorithms gives better results [22]. The one-class SVM technique for machine fault detection and classification in electro-mechanical machinery from vibrating measurements gives better performance in detecting outliers in comparison of multi-layered perception it is one of the artificial neural technique [23]. A new approach called Enhanced SVM approach is used for identifying attacks in network. It uses SOFM, PTF, GA, and packet-flow based data pre-processing [24]. A new feature selection method that increases the specificity and sensitivity, adding neural ensemble decision tree to evolve better optimal features increases the detection rate in comparison with various
methods such as Decision Stump, C4.5, Naive Bayes Tree, Random Forest, Random Tree and Respective tree model [25]. An integration of KNN and Unsupervised clustering algorithms for network anomaly detection, MLBG is used for reducing the time expense and increasing the performance in terms of accuracy in comparison to the un-weighted KNN classifier, using a genetic algorithm [26]. AFC-ANN solves the problems in the IDS. This approach achieves good detection precision rate and detection stability in comparison to the BPNN, Decision tree and Naive Bayes [28]. This method presents a combination of supervised tree classifier and unsupervised Bayesian clustering for IDS. This approach provides the high detection rate and false alarm rate in comparison of Kernel miner, Three-level tree classifier, Bagged boosted C5.0 trees [29].

3 Intelligent Ensemble Intrusion Detection Systems

3.1 NSL-KDD Dataset preparing and pre-processing

NSL-KDD Dataset was introduced by Tavallaee et al. [27], an enhanced version of KDDcup1999 benchmark dataset because of inherent issues. Statistical analysis conducted on KDDCup dataset found important issues that greatly affected the performance of anomaly detection system. NSL-KDD dataset contain 41 attributes and a label assigned to each assigned connection is either attack or normal. Pre-processing of NSL-KDD dataset first converts the continuous values into numeric values, data is cleaned and transformed following the filling up the missing values. The attributes are scaled and fall between 0 and 1.

3.1.2 K-Nearest Neighbourhood

KNN is simple and an effective technique for classification [1][13][26][21]. It assumes that the total sample set contains not only the data, but also preferred classification for each instances. When a classification is made to compute the sampling set its distance is to be computed and closest k instances is to be considered. The new instances are to be classified based in k closest instances. The Euclidean distance is used to measure the similarity of the instances. Let x1, x2, x3 and y1, y2, y3 are the features their Euclidean distance are measured as distance (X,Y) is

$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

After computing the Euclidean distance of all features the best quality of features is given to the C5.0 algorithm.

3.2 C5.0 Algorithm

C5.0 algorithm is the latest version of machine learning algorithms (MLAs) developed by Quinlan, built on decision tree [7]. The decision trees are constructed based on list of probable attributes and set of training instances, and then the tree can be classified by using following set of test instances. It is an advanced version of C4.5 classifier and it
has several important advantages over its ancestors [8]. C5.0 supports boosting and is a technique for integrating multiple classifiers to give improved final predictive accuracy. It incorporates variable misclassification costs [20].

C5.0 builds classifiers to diminish estimated misclassification costs rather than the fault rates. It supports sampling, cross-validation and quite robust in the presence of missing data. It does not require long training times to estimate. It has options to convert the tree to rules. C5.0 tree or rule sets are usually smaller than C4.5. Some other features of C5.0 are the soft or fuzzy thresholding which can also be specified. The asymmetric cost can be assigned to specific types of error. The confidence factor for pruning can also be changed. An option global pruning algorithm can be turned on/off. The minimum number of nodes in the terminal node can also be adjusted [9][20].

3.2.1 Information Gain and Entropy

Information gain is defined to decide how well a feature divides the training data agreeing to the target model. It is based on an amount commonly used in information theory known as entropy. The units of entropy are bits [22].

Let T is the training sample set.

\[ C_i = \text{Class I; } i = 1, 2, \ldots, n \]

\[ I(T_1, T_2, \ldots, T_n) = - \sum p_i \log_2 (p_i) \]

\[ T_i \] is the number of samples in class i

\[ P_i = T_i / T \]

\[ \log_2 \] is the binary Logarithm

Let attribute F have v distinct values

\[ \text{Entropy} = E(F) \] is

\[ \sum \left\{ \left( T_1 j + T_2 j + \ldots + T_n j \right) / T \right\} \ast I(T_1 j \ldots T_n) \quad j = 1 \]

Where \( T_i \) is Samples in Class i and subset j of attribute F

\[ I(T_1 j, T_2 j, \ldots, T_n j) = - \sum p_{ij} \log_2 (p_{ij}) \]

\[ Gain(F) = I(T_1, T_2, \ldots, T_n) - E(F) \]

3.3 One-class SVM

The One-class SVM was inspired by general SVM. One-class SVM is a famous outlier (or) novelty (or) anomaly detection algorithm in various application like machine fault detection and document classification [23]. It recognizes outliers among positive instances and remains as negative instances. It is used to classify abnormal activities as outliers. Let \( x_1, x_2, \ldots, x_l \in X \) be the training instances fitting to original space \( X \) and \( l \) be the number of
occurrences. It discovers the highest margin hyper plane that best divides the training instances from the origin. It is hard to discover a hyper plane that builds training data patterns separate from the origin in the original space $X$, the SVM uses a feature map ($\phi: X \rightarrow F$), which non-linearly transforms the instances from the original space to the feature space. The 1-class SVM is formulated using quadratic programming[20].

$$\min_{w,\varepsilon,\rho} \frac{1}{2}||w||^2 + \frac{1}{v} \sum_{i=1}^{l} \varepsilon_i - \rho \text{ Subject to } (w.\phi(x)) \geq \rho - \xi_i$$

$$\xi_i \geq 0, i = 1, ..., l$$

where, $w$ is the weight vector orthogonal to the hyperplane, $\varepsilon = (\varepsilon_1, ..., \varepsilon_l)$ is the vector of slack variable used to penalize the forbidden instances, and $\rho$ represents the margin, $v$ is the constraint that increases the distance of hyperplane from the origin and fraction data containing in the separate region. Due to curse of dimensionality[23][17], the SVM uses the kernel theory, the inner dot product in the feature space is calculated by kernel function $k(x, y) = \phi(x).\phi(y)$, such as Gaussian kernel, $k(x, y) = e^{-\gamma||x-y||^2}$. Using the kernel function and Lagrangian multiplier to the original quadratic programing Eq.3 builds a decision function. The test instance $(x)$ is

$$f(x) = \text{sgn}((w.\phi(x) + b) - \rho)$$

The test instance $(x)$ is accepted when $f(x)$ is positive and it is forbidden when $f(x)$ is negative. Positive instances shows that test instance $(x)$ is related to the training instances and the negative instances shows that it departs from the training data and is considered as an anomaly.

---

**Fig 1. Frame work of the Proposed System.**
C4.5 algorithm and one-class SVM implemented earlier achieves a detection rate of 88.5% and further proposed C5.0 and one-class SVM resulted in an increase to 95% [20]. Overall accuracy of 8% improvement is obtained. In order to achieve better feature selection, KNN is incorporated along with C5.0 and one-class SVM. A hybrid approach for better feature selection and anomaly detection with better classification results is proposed in order to increase the classification performance [20].

The proposed methodology of an intelligent intrusion detection system is shown in figure 1. NSL-KDD dataset is prepared and pre-processed. The pre-processed data is split as 20% of training and 20% of testing data. KNN is used for best optimal subset selection and to remove redundant and irrelevant features. The best quality subset feature is given to C5.0 algorithm and it is used to classify apart from sharing the node information. The collected information of anomaly data is given to one-class SVM, which further predicts the outliers from the classified information.

4. Conclusion

Intrusion detection is one of the main research challenges in computer security in order to detect unseen attacks. In present literature, procedures to detect anomaly is available. In order to achieve better performance, a hybrid approach with a combination of misuse detection and anomaly detection methods will give better results. An analysis on intelligent ensemble intrusion detection system based on k-nearest neighbourhood with C5.0 algorithm and one-class support vector machine model is proposed. Further, on analysis, we infer that combining KNN with C5.0 and One-class SVM for optimal subset selection and anomaly detection will provide improved performance.

References


20. Rani MS, Xavier SB. A Hybrid Intrusion Detection System Based on C5.0 Decision Tree and One-Class SVM.


