www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 1 January 2015, Page No. 10062-10064

An Intrusion Detection System Based On Support Vector Machine Using Hierarchical Clustering And Genetic Algorithm

Minakshi Bisen[1], Amit Dubey[2]

[1]Minakshi Bisen, M.Tech Scholar, department of Computer Science & Engineering,
Oriental College of Technology, Bhopal,
Madhya Pradesh, INDIA
E-Mail:er.mini27bisen.@gmail.com

[2]Amit Dubey,HOD department of Computer Science & Engineering,
Oriental College of Technology, Bhopal,
Madhya Pradesh, INDIA
E-Mail:amitdubey@oriental.ac.in

Abstract: An NIDS system based on SVM ,GA and Hierarchical Clustering is proposed.GA along with hierarchical clustering is used to provide fewer,efficient and abstraced instances of the KDD Cup 1999 dataset to SVM for further processing.GA is used to eliminate the unimportant feature and BIRCH hierarchical clustering is used to provide optimal instances of the data set to the SVM.Due to this optimal instances, SVM will be able to classify the network traffic data more accurately and precisely . This system try to reduce the detection time of r2l,u2r,probe and DoS atack and increase in the accuracy.

Keywords: Network Intrusion detection System ,Support vector machine, Hierarchical clustering, Genetic algorithm

1. Introduction

As the internet technology is growing day by day,lot of amount of e-commerce transaction are performed in a day.In the network environment ,Intrusion detection is the most important for security infrastructure. It is used to identify and detect the network traffic. Network intrusion detection system (NIDS), as an important link in the network security infrastructures, aims to detect malicious activities, such as denial of service attacks, port scans, or even attempts to crack into computers by monitoring network traffic. In addition to inspecting incoming network traffic, NIDS can also obtain valuable information on an ongoing intrusion from outgoing or local traffic. [15].

Nowadays,much attention has been paid to intrusion detection system (IDS) which is closely linked to the safe use of network services.However, it is not easy to discern the attack and the normal network visit. To overcome this problem, various artificial intelligence methods are developed, such as fuzzy logic [3] ,K-nearest neighbor [8], support vector machine, SVM[6], artificial neural networks, ANN [17], genetic algorithm [9] .

Among the methods mentioned above, SVM is an effective one,which is a well-known classifier tool based on small sample learning.Since SVM has manifested its robustness and efficiency in the network action classification, it therefore becomes a popular method widely used in IDS [16].This proposed system is designed in MATLAB R2012b(8.0.0.783).

2. Literature Survey

Intrusions pose a serious security threat for the stability and the security of information in a network environment. A network intrusion attack encompasses a wide range of activities. It includes attempting to destabilize the network, gaining unauthorized access to files with privileges, or mishandling and misusing of software. The intrusion detection is to automatically scan network activity and detect intrusion attacks [13]. The number of features extracted from raw network data, which an IDS needs to examine, is usually large even for a small network [9]. Many researchers have tried to improve the detection rate of IDS through proposing new classifiers, but improving the effectiveness of classifiers is not an easy task, Though feature selection can be used to optimize the existing classifiers. Feature selection is useful to reduce the computational complexity (reduce training and utilization times), remove information redundancy, increase the accuracy of the learning algorithm, facilitate data understanding and improve the generalization. In this line of research,some techniques have been used in developing a light weight IDS such as machine-learning technique to build mixed-probabilistic models from the training dataset and to determine if a given piece of data is an anomaly[18]. Some researchers has presented a clustering-based anomaly detection approach. Their approach creates clusters from the training dataset and automatically labels clusters as ‘normal’ or ‘anomalous’ in term of their sizes. It uses the labeled clusters to classify network data according to the label of the nearest cluster [5]. Many researchers have applied data mining techniques in the design of NIDS. One of the promising techniques is support vector machine (SVM), with solid mathematical foundations have provided satisfying results[7]. SVM separates data into multiple classes (at least two) by a hyperplane and simultaneously minimizes the empirical classification error and maximizes the geometric margin. Thus, it is also known as maximum margin classifiers [14]. The basic SVM deals with two-class problems—in which the data are separated by a hyperplane defined by a number of support vectors. Support vectors are a subset of training data used to define the boundary between the two classes. In situations where SVM cannot separate two classes, it solves this problem by mapping input data into high-dimensional feature spaces using a kernel function[17]. SCAN, a stochastic clustering method, which used expectation maximization to calculate the attribute value of the missing data, and also reduced the amount of data by feature selection[11]. A genetic algorithm (GA) for feature selection,and SVM for intrusion detection is used. BIRCH uses a hierarchical data structure called Clustering Feature tree (CF tree) for partitioning the incoming data points in an incremental and dynamic way. BIRCH stores fewer abstracted data points than the whole dataset. Each abstracted point represents the centroid of a cluster of data points[12].

3.Method Used

3.1. Support Vector Machine

Support vector machines (SVM) are a group of supervised learning methods that can be applied to classification or regression. It represents an extension to nonlinear models and is based on the statistical learning theory. It is simple,comes from the fact that Support Vector Machines apply a simple linear method to the data but in a high-dimensional feature space non-linearly related to the input space. Moreover, even though we can think of SVM as a linear algorithm in a high-dimensional space, in practice, it does not involve any computations in that high-dimensional space. This simplicity combined with state of the art performance on many learning problems (classification, regression, and novelty detection) has contributed to the popularity of the SVM.

Support Vector Machine is basically a classifier which performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables.

3.2 Genetic Algorithm

It is widely used for optimal feature selection.In the proposed methodology, GA provides,a framework to specify the parameters in an integrated context . To implement GA, the specification of an appropriate coding for each possible solution given. In this ,solution is defined by the attributes, which are used for model development and the parameter required to define the kernel function.GA is used to select the appropriate features from the large data set.

3.3 BIRCH: Balanced Iterative Reducing and Clustering Using Hierarchies

BIRCH is designed for clustering a large amount of numerical data. It integrates the hierarchical clustering (at the initial microclustering stage) and other clustering methods such Hierarchical Methods as iterative partitioning (at the later macroclustering stage).

It overcomes the two difficulties of agglomerative clustering methods:
(1) scalability and
(2) the inability to undo what was done in the previous step.

BIRCH introduces two concepts:
1.Clustering Feature (CF)
2.Clustering Feature tree (CF tree)

They are used to summarize cluster representations. These structures help the clustering method to achieve hierarchical methods, good speed and scalability in large databases and also make it effective for incremental and dynamic clustering of incoming objects.

Clustering features are additive. Clustering features are sufficient for calculating all of the measurements that are needed for making clustering decisions in BIRCH.

4. Result

In KDD Cup 1999 data set ,there are 4,898,431 and 311,029 records in training and test data set.In this data set, attack records were classified into four categories viz DOS,U2R,R2L and Probe. In the training set 19.85% were normal traffic and rest were attack traffic while for the test data set, it contains 19.48% as the normal traffic and rest were attack traffic.There are 41 quantitative and qualitative features in each record of KDD data set[15].

In other system , time to find out the various attack is more and also accuracy is not optimal.In this ,by using GA and hierarchical clustering, SVM is able to get prepocessed and optimal data-set,due to which detection time get reduce and accuracy is also get increases

Table 1: Comparison among accuracy,detection time and fp

CF tree	*Accuracy*	*Detection time*	*fp*
T=0.1	0.9667	0.024	0.069
T=0.2	0.9685	0.028	0.071
T=0.3	0.9677	0.022	0.067

5. Conclusion and Future Work

In this proposed work, an SVM along with Genetic algorithm and Hierarchical Clustering is used.An NIDS is designed with Genetic and BIRCH hierarchical clustering which provide preprocessed, reduced and abstracted data -set to SVM. Then , the SVM train this reduced data set and classify the attacks. SVM perform training on the reduced KDD Cup 1999 data set and detect the various attacks in less time and gives better accuracy and performance.In future, it can be performed with other data set and some new attack instances which are not detected in r2l and u2r attack can be detected.

6. References

[1] Alexandros Karatzoglou ,David Meyer Kurt ,Hornik Wirts chafts "Support Vector Machines in R" Journal of Statistical Software.

[2]. Ball, G., & Hall, D. (1967) "A clustering technique for summarizing multivariate data" Behavioral Science, 12, 153–155.

[3]. Chimphlee, W., Addullah, A. H., Sap, M. N. M., Srinoy, S., & Chimphlee, S. (2006) "Anomaly-based Intrusion detection using fuzzy rough clustering" In Paper presented at the international conference on hybrid information technology (ICHIT 06).

[4] Eskin, E., 2000. "Anomaly detection over noisy data using learned probability distributions." Proc. ICML, 255–262.

[5] Eskin, E., Arnold, A., Prerau, M., Portnoy, L., Stolfo, S., 2002 " A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In": Proc. Data Mining for Security Applications.

[6]. Joseph, J. F. C., Das, A., Lee, B. S., & Seet, B. C. (2010)." CARRADS: Cross layer based adaptive real-time routing attack detection system for MANETS " Computer Networks, 54(7), 1126–1141.

[7]. Khan, L., Awad, M., & Thuraisingham, B. (2007) "A new intrusion detection system using support vector machines and hierarchical clustering" The International Journal on Very Large Data Bases, 16(4), 507–521.

[8]. Li, Y., & Guo, L. (2007)" An active learning based TCM-KNN algorithm for supervised network intrusion detection", Computer and Security, 26, 459–467.

[9] Mukkamala, S., Sung, A. H., & Abraham, A. (2004) " Modeling intrusion detection systems using linear genetic programming approach. Proceedings of innovation in applied artificial intelligence" 17th international conference on industrial and engineering applications of artificial intelligence and expert systems (IEA/AIE). Lecture notes in computer science (Vol. 3029). Springer.

[10] Novikov D, Yampolskiy RV, Reznik L." Anomaly detection based intrusion detec- tion. In": Proceedings of the third international conference on information technology: new generations (ITNG'06); 2006.

[11] Patcha A, Park J-M. " An overview of anomaly detection techniques: existing solutions and latest technological trends". Computer Network 2007, doi:10.1016/j.comnet.2007.02.001.

[12] Shon, T., Kim, Y., Lee, C., & Moon, J. (2005). A machine learning framework for network anomaly detection using SVM and GA. In Proceedings of the 2005 IEEE workshop on information assurance and security (pp. 176–183).

[13] S. Jiang et al. "A clustering –based method for unsupervised intrusion detection " Pattern Recognition Letters 27 (2006) 802–810.

[14]. Shafi, K., & Abbass, H. A. (2009) " An adaptive genetic-based signature learning system for intrusion detection" Expert Systems with Applications, 36(10), 12036–12043.

[15]. S.-J. Horng et al. " A novel intrusion detection system based on hierarchical clustering and support vector machines ", Expert Systems with Applications 38 (2011) 306–313.

[16]. Tsai, C. F., Hsu, Y. F., Lin, C. Y., & Lin, W. Y. (2009) "Intrusion detection by machine learning ": A review. Expert Systems with Applications, 36, 11994–12000.

[17] Wang, G., Hao, J., Ma, J., & Huang, L. (2010)."A new approach to intrusion detection using artificial neural networks and fuzzy clustering" Expert Systems with Applications, 37, 6225–6232.

[18] Yamanishi, K., Takeuchi , J.I., 2001 "Discovering outlier filtering rules from unlabeled data: combining a supervised learner with an unsupervised learner In": Proc. ACM SIGKDD01, San Francisco, California, USA. 389–394.

Author Profile

Minakshi bisen was born in Bhopal in 1988.She received the BE degree (with distinction) in computer science and engineering from Sagar Institute Research and Technology bhopal, RGPV, Bhopal in 2011.She is currently pursuing M.Tech in computer science and engineering from Oriental College of Technology, RGPV, Bhopal**.** She has attended the national conference held in various institutes and presented papers in different research areas. Her research interests include data mining, network intrusion detection.