Anomaly network traffic detection using entropy calculation and support Vector machine

Basant Agarwal
Department of Computer Science & Engineering, Central University of Rajasthan, Kishangarh, Ajmer
thebasant@gmail.com

Abstract—*Intrusion detection systems (IDS) have a vital role in protecting computer networks and information systems. In this paper, we propose a method for identifying abnormal traffic behaviour based on entropy and support vector machine. Main challenge is to distinguish between normal traffic and attack traffic since there is no major difference between normal and attack traffic. Our objective is to extract network features and make a model to identify the attack traffic. We propose an anomaly network traffic detection method based on Support Vector Machine (SVM) and entropy of network parameters.*

Entropies of network parameters are extracted from the traffic coming in the network.Than Support vector machine model is developed to identify the attack traffic. The entropy of network traffic is calculated in certain duration, and then sends its outputs directly to the SVM model for analysis. We made two type of SVM model for identifying the attack traffic and normal traffic. Those are one class SVM and 2 dimensional SVM.

Experiments are performed on the 1999 DARPA Intrusion Detection Evaluation at Massachusetts Institute of Technology, Lincoln Lab. The first week of the data is attack free, while the second week of the data contains attacks. To evaluate the ability of the anomaly based intrusion detection system we only considering attack that has anomaly signature. Those are Portsweep, Ipsweep, Mailbomb, and Neptune. Experiment result demonstrates that our method works well with high detection rate of attack traffic and very less false alarm rate.

Keywords—*Intrusion detection, Denial of service attacks, Support vector machines, Entropy, Anomaly traffic detection.*

1. Introduction

Security of computers and the networks that connect them is increasingly becoming of great significance. Intrusion detection is a mechanism of providing security to computernetworks. Although there are some existing mechanisms for Intrusion detection, thereis need to improve the performance. Data mining techniques are a new approach for Intrusion detection. There is considerable interest in using entropy-based analysis of traffic feature distributionsfor anomaly detection. Entropy-based metrics are appealing since they provide more fine-grained insights into traffic structure than traditional traffic volume analysis.While previous work has demonstrated the benefits of using the entropy of different traffic distributions in isolation to detect anomalies, there has been little effort in comprehensivelyunderstanding the detection power provided by entropy-based analysis of multiple traffic distribution used in conjunction with each other.

Support vector machines (SVMs) are a set of related supervised learning methodsused for classification and regression. An SVM training algorithm builds a model thatpredicts whether a new example falls into one category or the other. Intuitively, an SVMmodel is a representation of the examples as points in space, mapped so that the examplesof the separate categories are divided by a clear gap that is as wide as possible. Newexamples are then mapped into that same space and predicted to belong to a categorybased on which side of the gap they fall on.

We have combined both entropy and support vector machines for the detectionof the anomaly based attacks. Our objective is to extract network features and make amodel to identify the attack traffic. We propose an anomaly network traffic detectionmethod based on Support Vector Machine (SVM) and entropy of network parameters.

First of all entropy of network parameters are extracted from the traffic coming in thenetwork than with these entropy values of network parameters we develop a supportvector machine model to identify the attack traffic.

Experiments are performed on the 1999 DARPA Intrusion Detection Evaluation atMassachusetts Institute of Technology, Lincoln Lab. The first week of the data is attackfree, while the second week of the data contains attacks. To evaluate the ability of theanomaly based intrusion detection system we only considering attack that has anomalysignature is count in counted. Those are Portsweep, Ipsweep, Mailbomb, and Neptune.Network parameters being extracted, are source IP address, destination IP address,source port address, destination port address, rate of occurrence

of UDP packets, rate ofoccurrence of TCP packets, rate of occurrence of ICMP packets, packet type and packetsize. The entropy of network traffic is calculated in certain duration, and then sends itsoutputs directly to the SVM for analysis. We made two type of SVM model for identifyingthe attack traffic and normal traffic. Those are one class SVM and 2 dimensional SVM.

The One-class SVM first models normal traffic distribution, then detects traffic that fallsoutside this distribution as attack traffic. The 2 dimensional SVM works with 2 classesone is attack traffic and other is normal traffic

2. Overview of Entropy calculation

Entropy is a measure of the uncertainty or randomness associated with a random variable or in this case data coming over the network. The more random it is, the more entropy it contains. The value of sample entropy lies in range [0, log (n)].The entropy shows its minimum value 0 when all the items (IP address or port) are same and its maximum value log (n) when all the items are different. The entropy of a random variable X with possible values {x1, x2, ...,xn} can be calculated as

$H(X) = -"P(x_i)\log P(x_i)$

Here, X for a fixed time window w is, $P(x_i) = m_i/m$, where m_i is the frequency or number of times we observe X taking the value x_i

$m = "m_i$

$H(X) = -"(m_i/m) \log (m_i/m)$

H (X) = Entropy

If we want to calculate probability of any source (destination) address then,

m_i = number of packets with x_i as source (Destination) address and

M = total number of packets

$P(x_i)$ =Number of packets with xi as source/destination address/M

Here total number of packets is the number of packets seen for a time window T.

Similarly we can calculate probability for each source (destination) port as

$P(x_i)$=Number of packets with x_i as source (destination) address/M

Normalized entropy calculates the over all probability distribution in the captured flow for the time window T.

Normalized entropy = (H/log(n))

Where n is the number of distinct x_i values in the given time window.

Entropy can be computed on a sample of consecutive packets. Comparing the value for entropy of some sample of packet header fields to that of other samples of packet header fields provides a mechanism for detecting changes in the randomness. And at the time of attack these values will deviate from its normal behaviour that we use to detect the anomaly attack.

Like we use entropy value of source IP address and destination IP address since that becomes small at the time of denial of service attack and increases the value of source and destination port numbers at the time of denial of service attack.In addition, the entropy value of packet type is worth observing because DDoS attacks use specific packet type such as ICMP flood attack and UDP flood attack. If the entropy of packet type converges to a small value, it needs to suspect to be under attack.

We can also use entropy of packet size of each packet. In attack situation, number of similar size packets should exceed than the number of types of packets in normal situation, hence entropy of packet size distribution should decrease in the attack situation.

3. Support Vector Machine

Support vector machine (SVM) is a non-linear classifier which is often reported as producing superior classification results compared to other methods. The idea behind the method is to non-linearly map the input data to some high dimensional space, where the data can be linearly separated, thus providing great classification (or regression) performance. A special property of SVM is that it simultaneously minimizes empirical classification errors and maximizes geometric margins. By training with DARPA dataset, SVM learns to find best compromise and gives the best projection results.

In SVM literature, a predictor variable is called anattribute, and a transformed attribute that is used to define the hyperplane is called a feature. The task of choosing the most suitable representation is known as feature selection. A set of features that describes one case (i.e., a row of predictor values) is called a vector. So the goal of SVM modelling is to find the optimal hyperplane that separates clusters of vector in such a way that cases with one category of the target variable are on one side of the plane and cases with the other category are on the other size of the plane. The vectors near the hyperplane are the support vectors.

A Two-Dimensional SVM Example

We wish to perform a classification, and our data has a categorical target variable with two categories. Also assume that there are two predictor variables with continuous values. If we plot the data points using the value of one predictor on the X axis and the other on the Y axis we might end up with an image such as shown below. One category of the target variable is represented by rectangles while the other category is represented by ovals.

In this example, the cases with one category are in the lower left corner and the cases with the other category are in the upper right corner; the cases are completely separated.

The SVM analysis attempts to find a 1-dimensional hyperplane (i.e. a line) that separates the cases based on their target categories. There are an infinite number of possible lines; two possible lines are shown above. The question is which line is better, and how do we define the optimal line. The dashed lines drawn parallel to the separating line mark the distance between the dividing line and the closest vectors to the line. The distance between the dashed lines is called the margin. The vectors (points) that constrain the width of the margin are the support vectors. The Figure 3.1 illustrates this

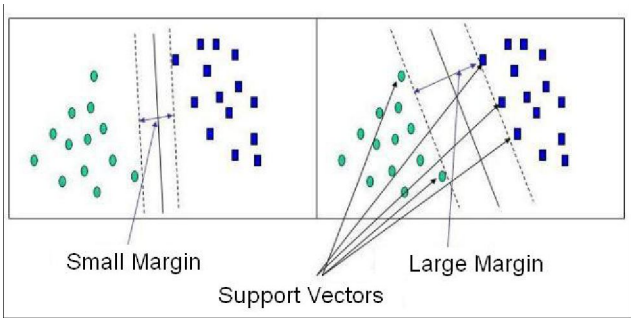

Figure 3.1: Support Vector Machine

An SVM analysis finds the line (or, in general, hyperplane) that is oriented so that the margin between the support vectors is maximized. In the figure 3.1, the line in the right panel is superior to the line in the left panel.

In some non-linear cases, SVM uses a kernel function to map the data into a different space where a hyperplane can be used to do the separation. The kernel function may transform the data into a higher dimensional space to make it possible to perform the separation. Some of kernel functions are linear, polynomial, radial

basis function, sigmoid. Ideally an SVM analysis should produce a hyperplane that completely separates the feature vectors into two non-overlapping groups. However, perfect separation may not be possible, or it may result in a model with so many feature vector dimensions that the model does not generalize well to other data; this is known as over fitting.

To allow some flexibility in separating the categories, SVM models have a cost parameter C that controls the trade of between allowing training errors and forcing rigid margins. It creates a soft margin that permits some misclassifications. Increasing the value of C increases the cost of misclassifying points and forces the creation of a more accurate model that may not generalize well.

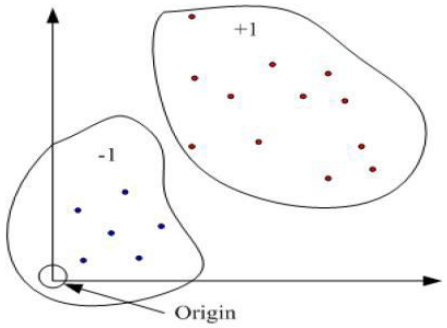

Figure 3.2 One-class SVM

One-class Support Vector Machine

One-class Support Vector Machine (SVM) was proposed by Scholkopf for estimating the support of a high-dimensional distribution [2]. Given a training dataset without any class information, the One-class SVM constructs a decision function that takes the value +1 in a small region capturing most of the data points, and -1 elsewhere. The strategy in this technique is to map the input vectors into a high dimension feature space corresponding to a kernel, and construct a linear decision function in this space to separate the dataset from the origin with maximum margin. Via the freedom to utilize different types of kernel, the linear decision functions in the feature space are equivalent to a variety of non-linear decision functions in the input space. The One-class SVM introduces a parameter v ϵ (0, 1) to control a tradeoff between the fractions of data points in the region and the generalization ability of the decision function. One-class SVM is represented in figure 3.2.

4. Experimental Setup

4.1 Data Source

Experiments are performed on the 1999 DARPA Intrusion Detection Evaluation at MassachusettsInstitute of Technology, Lincoln Lab (http://www.ll.mit.edu/IST/ideval/). The first week of the data is attack free, while the second week of the data containsattacks. So that, we use the inside-tcpdump data set in the first week for training andin the second week for testing. To evaluate the ability of the anomaly based intrusiondetection system we are only considering attack that has anomaly signature. The detailsof those attacks are shown in table 4.1.

- Portsweep, surveillance sweep through many ports to determine which services are supported on a single host. Network traffic may contain anomaly number of TCP,UDP and ICMP packets.
- Mailbomb, a denial of service attack where we send the mail server many large messages for delivery in order to slow it down. Network traffic may contain anomalynumber of SMTP packets.
- Ipsweep, surveillance sweep performing either a port sweep or ping on multiple host addresses. Network traffic may contain anomaly number of ICMP packets.
- Satan, a network probing tool which looks for well-known weaknesses. Network traffic may contain anomaly number of TCP and UDP packets.
- Neptune, SYN flood denial of service on one or more ports. Network traffic maycontains anomaly number of SYN and RET packets.

Day	Name	Destination	Start Time	End Time
Tue	Portsweep	172.16.114.50	08:44:15	09:11:11
	Mailbomb	172.16.112.50	14:25:10	14:35:06
	Ipsweep	172.16.112.0/23	13:05:13	13:29:14
Wed	Satan	172.16.114.50	12:02:09	12:04:38
	Mailbomb	172.16.112.50	13:44:13	13:54:08
	Ipsweep	172.16.112.0/23	20:17:04	20:29:13
Thur	Satan	172.16.114.50	09:33:20	09:35:37
	Portsweep	172.16.114.50	10:50:07	11:07:31
	Neptune	172.16.114.207	11:04:12	11:04:37
	Ipsweep	172.16.112.0/23	16:36:06	16:36:33
Fri	Neptune	172.16.114.50	11:20:11	11:23:36
	Portsweep	172.16.112.50	17:13:02	17:25:06z

4.2 Proposed Procedure

The procedure of making a support vector machine model is as following and is further explained in subsections.

1. Extract features from the dataset.
2. Transform data to the format of an SVM package.
3. Conduct simple scaling on the data.
4. Consider the RBF kernel.
5. Find the best parameter C and γ .
6. Use the best parameter C and γ to train the whole training set.
7. Test.

4.2.1 Feature extraction

Entropy-based analysis of traffic feature distributions is used for anomaly detection. Entropy-based metrics are appealing since they provide more fine-grained insights intotraffic structure than traditional traffic volume analysis.Features extracted for the detection of anomaly based attack are as follows.

- Entropy of source IP address and port number.
- Entropy of destination IP address and port number.
- Entropy of packet type (ICMP, TCP, UDP).
- Occurrence rate of packet type (ICMP, UDP, TCP).
- Number of packets per unit time.
- Entropy of packet size.

4.2.2 Transform data to the format of an SVM package

LibSVM[4] is used for making support vector machine model. We wrote program in java to convert data in the input format for libSVM. We conducted scaling on thedata so that all the data is in the same range. The main advantage of scaling is toavoid attributes in greater numeric ranges dominating those in smaller numeric ranges.Another advantage is to avoid numerical difficulties during the calculation.

4.2.3 Model selection

4.2.3.1 Model selection for two-class support vector machine

There are two parameters for an RBF kernel: C and . It is not known beforehand which C and is best for a given problem; consequently some kind of model selection(parameter search) must be done. The goal is to identify good (C,) so that the classifiercan accurately predict unknown data (i.e. testing data).Common strategy is to separate the data set into two parts, of which one is considered unknown. The prediction accuracy obtained from the unknown set more precisely reflects. The performance on classifying an independent data set. An improved version of thisprocedure is known as cross-validation.

In v-fold cross-validation, we first divide the training set into v subsets of equal size.Sequentially one subset is tested using the classifier trained on the remaining v-1 subsets.Thus, each instance of the whole training set is predicted once so the cross-validationaccuracy is the percentage of data which are correctly classified. The cross-validation procedure can prevent the over fitting problem. The best method is to use grid-searchon C and using cross-validation. Various pairs of (C,) values are tried and the onewith the best cross-validation accuracy is picked.

4.2.3.2 Model selection for one-class support vector machine

There are two parameters need to be set before training the One-class SVM: v and. The generalization performance of one-class SVM can be evaluated by two measures:the size of region and the generalization fraction of data points in the region. Small sizeindicates the probability that a data point of class “-1” falls into this region is small.Great generalization fraction of data points indicates that the probability that a newdata point of this class (+1) falls into this region is great. The parameter decides thenon-linear characteristics of the decision function, in other words, it decides the “shape” of the region. The parameter controls not only the fraction of data points in the region but also the generalization ability. Thus, the kernel and v all influence the size of region and the generalization fraction of data points in the region. In a word, they influence the generalization performance of one-class SVM. We tried different combinations of v and to get the best results.

4.2.4 Results

Our experiments have two phases, namely training and a testing phase. In the training phase the system constructs a model using the training data. The test data is testedwith the constructed model to detect the intrusion in the testing phase. In order to havea good prediction performance, IDS should be able to correctly differentiate betweenintrusions and legitimate actions in a system environment. We use standard measurements such as detection rate (DR), false positive rate (FPR) and overall classification rates (CR) to evaluate the performance of intrusion detection tasks.

- True Positive (TP): The number of malicious records that are correctly identified.
- True Negatives (TN): The number of legitimate records that are correctly classified.
- False Positive (FP): The number of records that were incorrectly identified as attackshowever in fact they are legitimate activities.
- False Negative (FN): The number of records that were incorrectly classified aslegitimate activities however in fact they are malicious.

DR =TP/ (TP + FN)

FPR =FP/ (TN + FP)

CR = (TP + TN)/ (TP + TN + FP + FN)

Actual Normal Connection	True Negative	False Positive
	85.71%	14.29%
Actual Intrusions (Attacks)	False Negative	True Positive
	1.88%	98.12%

Table 4.2 Predicted Normal Predicted Intrusions for 2 class SVM

Actual Normal Connection	True Negative	False Positive
	82.12%	17.88%
Actual Intrusions (Attacks)	False Negative	True Positive
	3.88%	96.12%

Table 4.3 Predicted Normal Predicted Intrusions for one-class SVM

Our experimental results for the 2 class Support vector machine are as follows: Truenegative rate correctly recognizing the normal connection is 85.71%. True positive ratecorrectly determining intrusions is 98.12%. While experimental results for the two -classSVM and one-class support vector machine are given in table 4.2 and 4.3. For one classSVM true positive rate correctly determining intrusions is 96.12 % and true negative ratecorrectly recognizing the normal connections is 82.12%. We used LibSVM for makingsupport vector machine model.

5. Conclusion and Future Work

We use SVM on the learning-based anomaly detection system, whereas in the choiceof tools, we use LibSVM as a SVM tool. We obtained good result with average detectionrate up to 98% for two class support vector machine while 96% for the one class supportvector machine by using LibSVM, with kernel (RBF), and without the need of other external kernels. As for two-class and one-class SVM, the result of detection rate andfalse positive are very good. However as results for two-class SVM are better than oneclass SVM but one class SVM is easier to model than two class SVM. Our method issimple and effective that it can achieve high detection rates.We have obtained nice results by using LibSVM with the DARPA dataset 1999 andtwo forms of SVM. But since both the attacking technology and the detection technologyneed to keep updating very fast, our future work is to design new methods and trainthem with latest data, and to classify data efficiently with new forms of SVM.

6. References

[1] Tan, P.-N., Steinbach, M., and Kumar, V. 2005. Introduction to Data Mining.Addison-Wesley.

[2] SchÃolkopf, B., Platt, J. C., Shawe-Taylor, J. C., Smola, A. J., and Williamson, R.C. 2001. Estimating the support of a high-dimensional distribution.Neural Comput.13, 7, 1443-1471

[3] O. Depren, M. Topallar, E. Anarim and M. K. Ciliz, An Intelligent Intrusion DetectionSystem (IDS) for Anomaly and Misuse Detection in Computer Networks.Expert Systems with Applications, vol. 29, pp. 713-722, 2005.

[4] Guide of LIBSVM. http://www.csie.ntu.edu.tw/ cjlin/papers/guide/guide.pdf

[5] Stefano, C., Sansone, C., and Vento, M. 2000. To reject or not to reject: that isthe question an answer in case of neural classifiers. IEEE Transactions on Systems,Management and Cybernetics 30, 1, 84-94.

[6] Barbara, D., Couto, J., Jajodia, S., and Wu, N. 2001b. Detecting novel networkintrusions using bayes estimators. In Proceedings of the First SIAM InternationalConference on Data Mining.

[7] Barbara, D., Li, Y., Couto, J., Lin, J.-L., and Jajodia, S. 2003. Bootstrapping a datamining intrusion detection system. In Proceedings of the 2003 ACM symposium onApplied computing. ACM Press, 421-425.

[8] Vapnik, V. N. 1995. The nature of statistical learning theory. Springer-Verlag NewYork, Inc., New York, NY, USA.

[9] Tax, D. and Duin, R. 1999a. Data domain description using support vectors.In Proceedingsof the European Symposium on Arti_cial Neural Networks, M. Verleysen,Ed. Brussels, 251-256.

[10] Tan, P.-N., Steinbach, M., and Kumar, V. 2005. Introduction to Data Mining.Addison-Wesley

[11] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. 1996. A density-based algorithm fordiscovering clusters in large spatial databases with noise. In Proceedings of SecondInternational Conference on Knowledge Discovery and Data Mining, E. Simoudis, J.Han, and U. Fayyad, Eds. AAAI Press, Portland, Oregon, 226-231.

[12] Guha, S., Rastogi, R., and Shim, K. 2000. ROCK: A robust clustering algorithm forcategorical attributes. Information Systems 25, 5, 345-366.

[13] L. Feinstein, D. Schnackenberg, R. Balupari, and D. Kindred. “Statistical approachesto DDoS attack detection and response”. In Proc. of DARPA Information SurvivabilityConference and Exposition (DISCEX 2003), pp. 303-314, April 2003.

[14] A. Lakhina, M. Crovella and C. Diot, “Mining anomalies using traffic feature Distributions”.In proc. of ACM SIGCOMM, pp. 217-228. August 2005.

[15] W. Lee, D. Xiang, “Information-theoretic measures for anomaly Detection”, In Proc.of IEEE Symposium on Security and Privacy, (Oakland, CA), pp. 130-143, 2001.

[16] Nychis, G., Sekar, V., Andersen, D. G., Kim, H., and Zhang, H. "An EmpiricalEvaluation of Entropy-Based Tra_c Anomaly Detection". Tech. Rep. CMU-CS-08-145, Computer Science Department, Carnegie Mellon University, 2008.

[17] Wagner, A., and Plattner, B. "Entropy Based Worm and Anomaly Detection in FastIP Networks".In Proc. IEEE WETICE (2005).

[18] K. Kumar, R.C., Joshi, and K., Singh, "A Distributed Approach using Entropyto Detect DDoS Attacks in ISP Domain," the International Conference on SignalProcessing, Communications and Networking, 2007. CSCN '07. Feb. 2007 pp:331 -337

[19] DARPA Intrusion Detection Evaluation,http://www.ll.mit.edu/IST/ideval/docs/docs_index.html, 1999.

[20] Ping Du ,Shunji Abe , "Detecting DoS Attacks Using Packet Size Distribution"BioneticsDec. 2007, p- 93-96.

[21] Keunsoo Lee, Juhyun Kim, Ki Hoon Kwon, Younggoo Han, Sehun Kim "DDoSattack detection method using cluster analysis" Expert Systems with Applications:An International Journal Volume 34, Issue 3 (April 2008)Pages 1659-1665.

[22] Sumitkar, Bibhudattasahoo , "An Anomaly detection system for ddos attack in gridcomputing" international journal of computer applications in engineering, technologyand sciences (IJ-CA-ETS) April '09 â September' 09 , Volume 1 : Issue 2 , Page:553-557.