CYBER THREAT ANALYSIS – A KEY ENABLING TECHNOLOGY FOR THE OBJECTIVE FORCE (A CASE STUDY IN NETWORK INTRUSION DETECTION)

Aleksandar Lazarevic*, Paul Dokas, Levent Ertoz, Vipin Kumar, Jaideep Srivastava, Pang-Ning Tan
Army High Performance Computing Reearch Center, Computer Science Department, University of Minnesota
aleks@cs.umn.edu, srivasta@cs.umn.edu, kumar@cs.umn.edu

ABSTRACT

The information technology advances that provide new capabilities to our forces also provide the enemy with new and powerful tools to launch attacks on our critical information resources. A specific example of this trend is the rapidly increasing rate of cyber attacks against our computers in the past few years. Traditional signature-based intrusion detection systems can only detect cyber attacks with known signatures. Our research focuses on applying data mining to build rare class prediction models for identifying known intrusions and their variations, as well as anomaly/outlier detection schemes for detecting novel attacks whose nature is unknown. Experimental results on the KDDCup'99 data set have demonstrated that our rare class predictive models are much more efficient in the detection of intrusive behavior than standard classification techniques. Experimental results on the DARPA 1998 data set, as well as on live network traffic at the University of Minnesota, show that the new techniques show great promise in detecting novel intrusions. In particular, during the past few months our techniques have been successful in automatically identifying several novel intrusions that could not be detected using state-of-the-art tools such as SNORT. In fact, many of these have been on the CERT/CC list of recent advisories and incident notes.

1. INTRODUCTION

Today computers control power, oil and gas delivery, communication systems, transportation networks, banking and financial services, and various other infrastructure services critical to the functioning of our society. Since the Second World War, science and technology have been a key enabler of the US military's global leadership. Progress in information technology is the critical to the ongoing transformation and eventual fielding of the Objective Force, as spelled out by many of the service's leaders at the Association of the U.S. Army's 2002 Winter Symposium. According to the (US Army White Paper 2001) "Concepts for the Objective Force", soldiers and leaders enabled by advanced technologies will provide revolutionary increases in operational capability. In addition, information systems will provide dominant situational understanding, enabling combined arms units to conduct simultaneous, non-contiguous, distributed operations. Platform designs in an arrangement of system-of-systems technologies will enable decisive maneuver, both horizontal and vertical, during day and night, and in all terrain and weather conditions. These breakthroughs will give Objective Force units the lethality and survivability needed to deliver full spectrum dominance, the versatility to change patterns of operation faster than the enemy can respond, and the agility to adjust to enemy changes of operation faster than he can exploit them (US white 2001).

Despite the tremendous benefits that information technology brings, there is also an escalation of the "dark side of the force" in the form of cyber terrorism, which is the use of informational technology capabilities to launch an attack on an organization's information resources. Today our real assets are stored electronically, and the targets are increasingly not only government and military installations, but public and private computer network systems as well. Information warfare extends the battlefield to incorporate all aspects of society given that cyber attacks have no boundary. To compound the problem, military and law enforcement authorities report that every month, assailants make thousands of unauthorized attempts to gain access to these systems, amounting to a nearly continuous assault (Vizard 1999).

2. CYBER THREAT ANALYSIS

According to a recent survey (Riptech 2001) by CERT/CC (Computer Emergency Response Team/Coordination Center), the rate of cyber attacks has been more than doubling every year in recent times – see Fig. 1. It has become increasingly important to make our information systems – especially those used for critical functions in the military and commercial sectors – resistant to and tolerant of such attacks. The key question is whether contemporary information technologies such as data mining can contribute to this battle and further enhance defense mechanisms. This paper addresses some possible directions in this battle.

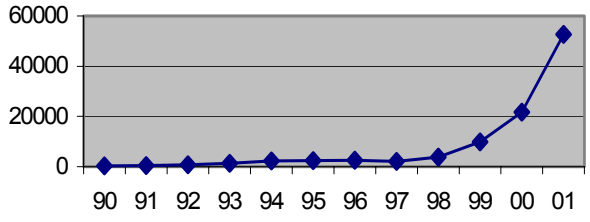

Fig.1. Cyber Incidents Reported to CERT/CC

With an eye towards the future, the Army is undergoing a transformation from a forward deployed 'Cold War' army to a power projection force. This transition will eventually result in a fully digitized, more configurable, rapidly expandable, strategically deployable, and effectively employable organization. It is clearly evident that the advanced information technologies will play an important role in this transition. Cyber Threat Analysis, as one of the emerging advanced technologies, has many different components including information assurance, methods to identify the most critical infrastructures, methods to detect cyber terrorist attacks and protect against cyber terrorism, intrusion detection and recovery from intrusions. All these components may affect army doctrines, tactics, techniques, and procedures on how we integrate digitized and non-digitized systems and organizations into the fight.

3. NETWORK INTRUSION DETECTION

The most widely deployed methods for detecting cyber attacks that employ signature-based detection techniques can only detect previously known attacks, since the signature database has to be manually revised for each new type of attack that is discovered. These limitations have led to an increasing interest in intrusion detection techniques based on data mining (Lee 1998, Barbara 2001) that generally fall into one of two categories; namely misuse detection and anomaly detection. In misuse detection, each instance in a data set is labeled as normal or intrusion (attack) and a learning algorithm is trained over the labeled data. These approaches are able to automatically retrain intrusion detection models on different input data that include new types of attacks as long as they have been labeled appropriately. Unlike signature-based intrusion detection systems, in misuse detection there is no need to manually construct the signatures, since automatically created models will capture the behavior of new types of attacks. Since such defined new behavior does not necessary describe only the single attack, but also its variations, misuse detection models are potentially more precise than manually created signatures. A key advantage of misuse detection techniques is their high degree of accuracy in detecting known attacks and their variations. Their obvious drawback is the inability to detect attacks whose instances have not yet been observed and whose behavior is significantly different than for the observed ones.

Traditional anomaly detection approaches, on the other hand, build models of normal data and detect deviations from the normal model in observed data. Anomaly detection algorithms have the advantage that they do not require the information about the attacks and can detect new types of intrusions as deviations from normal usage (Denning 1987). In this problem, given a set of normal data to train from, and given a new piece of test data, the goal of the intrusion detection algorithm is to determine whether the test data belong to “normal” or to an anomalous behavior. However, anomaly detection schemes suffer from a high rate of false alarms. This occurs primarily because previously unseen (yet legitimate) system behaviors are also recognized as anomalies, and hence flagged as potential intrusions.

This paper presents the scope and status of our research work both in misuse detection and anomaly detection. After the brief overview of our research in building predictive models for learning from rare classes, the paper gives a comparative study of several anomaly detection schemes for identifying novel network intrusions. We present experimental results on DARPA 1998 Intrusion Detection Evaluation Data, the KDDCup’99 data set, as well as on real network data from the University of Minnesota. Experimental results on the KDDCup’99 data set have demonstrated that our rare class predictive models are much more efficient in the detection of intrusive behavior than standard classification techniques. Experimental results on the DARPA 1998 data set [9], as well as on live network traffic at the University of Minnesota, show that the new techniques show great promise in detecting novel intrusions. In particular, during the past few months our techniques have been successful in automatically identifying several novel intrusions that could not be detected using state-of-the-art tools such as SNORT. In fact, many of these have been on the CERT/CC list of recent advisories and incident notes.

4. LEARNING FROM RARE CLASSES

In misuse detection related problems, standard data mining techniques are not applicable due to several specific details that include dealing with skewed class distribution, learning from data streams (intrusions are sequences of events) and proper labeling network connections. The problem of skewed class distribution is very pronounced in the network intrusion detection since intrusion as a class of interest is much smaller i.e. rarer than the class representing normal network behavior. In such scenarios when the normal behavior may typically represent 98-99% of the entire population a trivial classifier that labels everything with the majority class can achieve 98-99% accuracy. It is apparent that in this case classification accuracy is not sufficient as a standard performance measure. ROC analysis and metrics such as *precision*, *recall* and *F-value* (Joshi 2001, 2002) have been used to understand the performance of the learning algorithm on the minority class. A confusion matrix, shown in Table 1, is used to define these metrics.

From Table 1, *recall*, *precision* and *F-value* may be defined as follows:

$$Precision = TP / (TP + FP)$$
$$Recall = TP / (TP + FN)$$
$$F\text{-}value = \frac{(1+\beta^2)\cdot Recall \cdot Precision}{\beta^2 \cdot Recall + Precision},$$

where β corresponds to relative importance of *precision* vs. *recall* and is usually set to 1.

Table 1. Standard metrics for evaluations of intrusions

Confusion matrix (Standard metrics)		Predicted connection label	
		Normal	Intrusions(Attacks)
Actual connection label	Normal	True Negative (TN)	False Alarm (FP)
	Intrusions (Attacks)	False Negative (FN)	Correctly detected attacks (TP)

We have developed several novel classification algorithms designed especially for learning from the rare classes. For example, PN rule (Joshi 2001b) is a two-stage learning algorithm based on computing the rules. The first stage is aimed at discovering P-rules that cover most of the intrusive examples, while the second stage discovers N-rules and eliminates false alarms generated in the first phase. CREDOS (Joshi) is a novel algorithm that first uses the ripple down rules to overfit the training data and then to prune them to improve generalization capability.

In data mining community it is well known that a combination of classifiers can be an effective technique for improving prediction accuracy. Rare-Boost (Joshi 2001, 2002) attempts to incorporate rare class learning algorithms into the boosting technique. Unlike standard boosting technique where the weights of the examples are updated uniformly, in Rare-Boost the weights are updated differently for all four entries shown in Table 1. This paper shows that our algorithms for learning from rare class when integrated within the boosting algorithm produce significantly better performance regarding better recall/precision balance than the boosting technique applied on standard data mining algorithms. SMOTEBoost (Lazarevic 2002) further investigates this idea by embedding the procedure for generating artificial examples from the minority (intrusion) class within the boosting procedure. Artificial examples are created after each boosting round, classifiers are then built on such newly generated data and finally they are combined using the boosting technique.

We have also investigated a standard association-based classification algorithm in order to focus on a rare class problem. First, a frequent itemset generation algorithm is applied to each class and then the best itemsets are selected as "meta-features". These constructed features are added to the original data set and a standard classification algorithm is applied to such obtained data set. Current classification algorithms based on associations use confidence-like measures to select the best rules to be added as features into the classifiers. However, these techniques may work well only if each class is well-represented in the data set. For the rare class problems, some of the high recall itemsets could be also beneficial as long as their precision is not too low. Therefore, the best itemsets that will be added to the original data set are selected not only according to the precision but also according to high recall and F-value.

5. ANOMALY DETECTION ALGORITHMS

For the case where the nature of the attack is unknown, we have developed anomaly and outlier detection schemes to detect novel attacks/anomalies. Most anomaly detection algorithms require a set of purely normal data to train the model, and they implicitly assume that anomalies can be treated as patterns not observed before. Since an outlier may be defined as a data point which is very different from the rest of the data, based on some measure, we employ several outlier detection schemes (Lazarevic) in order to see how efficiently these schemes may deal with the problem of anomaly detection.

Nearest Neighbor (NN) Approach. This approach is based on the distance $D^k(O)$ of the k-th nearest neighbor from the point O. For instance, points with larger values $D^k(O)$ have more sparse neighborhoods and they typically represent stronger outliers than points belonging to dense clusters. In our *NN approach* we chose $k = 1$ and specify an "outlier threshold" that will serve to determine whether the point is an outlier or not. The threshold is based only on the training data and it is set to 2%. In order to compute the threshold, for all data points from training data (e.g. "normal behavior") distances to their nearest neighbors are computed and then sorted. All test data points that have distances to their nearest neighbors greater than the threshold are detected as outliers.

Mahalanobis-distance Based Outlier Detection. Since the training data corresponds to "normal behavior", the Mahalanobis distance (Mahalanobis 1930) is computed between the particular point p and the mean μ of the normal data. Similarly to the previous approach, the threshold is computed according to the most distant points from the mean of the "normal" data and it is set to be 2% of total number of points. All test data points that have distances to the mean of the training "normal" data greater than the threshold are detected as outliers. Computing distances using standard Euclidean distance metric is not always beneficial, especially when the data has a distribution similar to that presented in Fig. 2. When using standard Euclidean metric, the distance between p_2 and its nearest neighbor is greater than the distance from p_1 to its nearest neighbor. However, when using the Mahalanobis metric, these two distances are the same. It is apparent

that in these scenarios, Mahalanobis based approach is beneficial compared to the Euclidean metric.

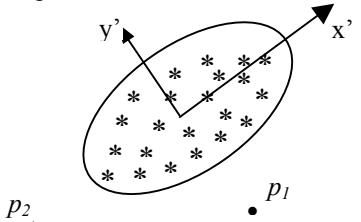

Fig. 2. Advantage of Mahalanobis-based approach

Density Based Local Outliers (LOF approach). The key idea of this method (Breunig 2000) is to assign to each data example a degree of being outlier, which is called the *local outlier factor (LOF)*. The outlier factor is local in the sense that only a restricted neighborhood of each object is considered. For each data example, the density of the neighborhood is first computed. The *LOF* of specific data example *p* represents the average of the ratios of the density of the example *p* and the density of its nearest neighbors. To illustrate advantages of the *LOF approach*, consider a simple two-dimensional data set given in Fig. 3. It is apparent that there is much larger number of examples in the cluster C_1 than in the cluster C_2, and that the density of the cluster C_2 is significantly higher that the density of the cluster C_1. Due to the low density of the cluster C_1 for every example *q* inside the cluster C_1, the distance between the example *q* and its nearest neighbor is evidently greater than the distance between the example p_2 and its nearest neighbor which is from the cluster C_2, and therefore example p_2 will not be considered as outlier. Therefore, the simple *NN approach* based on computing the distances fail in these scenarios. However, the example p_1 may be detected as outlier using the distances to the nearest neighbor. On the other side, *LOF* is able to capture both outliers (p_1 and p_2) due to the fact that it considers the density around the points.

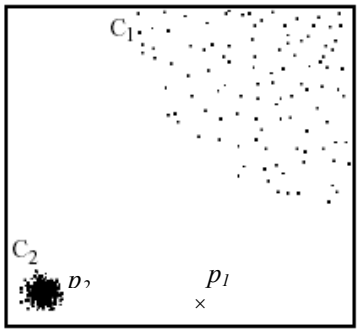

Fig.3. Advantages of the LOF approach

Unsupervised Support Vector Machines. Unlike standard supervised support vector machines (SVMs) that require labeled training data to create their classification rule, in (Schölkopf 2001) the SVM algorithm was adapted into unsupervised learning algorithm. This modification does not require training data to be labeled to determine a decision surface. Whereas the supervised SVM algorithm tries to maximally separate two classes of data in feature space by a hyperplane, the unsupervised algorithm attempts to separate the entire set of training data from the origin, i.e. to find a small region where most of the data lies and label data points in this region as one class. Points in other regions are labeled as another class. By using different values for SVM parameters (variance parameter of radial basis functions (RBFs), expected outlier rate), the models with different complexity may be built. For RBF kernels with smaller variance, the number of support vectors is larger and the decision boundaries are more complex, thus resulting in very high detection rate but very high false alarm rate too. On the other hand, by considering RBF kernels with larger variance, the number of support vectors decreases while the boundary regions become more general, which results in lower detection rate but lower false alarm rate too.

6. EXPERIMENTS

Our experiments are first performed on 1998 DARPA Intrusion Detection Evaluation Data (Lippmann 2000) and on its modification, KDDCup'99 data set (Lee 1998). The DARPA'98 data contains both training data and test data. The training data consists of 7 weeks of labeled network-based attacks inserted in the normal background data. The test data contained 2 weeks of network-based attacks and normal background data. The data contains four main categories of attacks:

- DoS (Denial of Service), for example, ping-of-death, teardrop, smurf, SYN flood, etc.,
- R2L, unauthorized access from a remote machine, for example, guessing password,
- U2R, unauthorized access to local superuser privileges by a local unprivileged user, for example, various buffer overflow attacks,
- PROBING, surveillance and probing, for example, port-scan, ping-sweep, etc.

Although DARPA'98 evaluation represents a significant contribution to the field of intrusion detection, there are many unresolved issues associated with its design and execution. In his critique, (McHugh 2000) questioned a number of results of DARPA evaluation, starting from usage of synthetic simulated data for the background (normal data) and using attacks implemented via scripts and programs collected from a variety of sources. In addition, it is known that the background data contains none of the background noise (packet storms, strange fragments, etc.) that characterizes real data. However, in the lack of better benchmarks, vast amount of the research is based on the experiments performed on this data set and its modification, KDDCup'99 data. However, in order to assess the performance of our anomaly detection algorithms in a real setting, we also applied our techniques to real network data from the University of Minnesota.

6.1. Feature construction

We used *tcptrace* utility software (tcptrace software tool) as the packet filtering tool in order to extract information about packets from TCP connections and to construct new features. The DARPA98 training data includes "list files" that identify the time stamps (start time and duration), service type, source IP address, source port, destination IP address, destination port and the type of each attack. We used this information to map the connection records from "list files" to the connections obtained using *tcptrace* utility software and to correctly label each connection record with "normal" or an attack type. A similar technique was used to construct KDDCup'99 data set (Lee 1998), but this data set did not keep the time information about the attacks. Therefore, we constructed our own features that were very similar in nature. These features include the number of packets, data bytes, acknowledgment packets, retransmitted packets, pushed packets, SYN and FIN packets flowing from source to destination as well as from destination to source. We have also added connection status as the content-based feature. The main reason for this procedure is to associate new constructed features with the connection records from "list files" and to create more informative data set for learning. However, this procedure was applied only to TCP connection records, since *tcptrace* software utility was not able to handle ICMP and UDP packets. For these connection records, in addition to the features provided by DARPA, we used the features that represented the number of packets that flowed from source to destination.

Since majority of the DoS and probing attacks may use hundreds of packets or connections, we have constructed time-based features that attempt to capture previous recent connections with similar characteristics. The same approach was used for constructing features in KDDCup'99 data (Lee 1998), but our own features examine only the connection records in the past 5 seconds. These features include the number of connections by the same source or to the same destination in last 5 seconds, and the number of different services from the same source or to the same destination as the current record in the last 5 seconds. "Slow" probing attacks that scan the hosts (or ports) and use a much larger interval than 5 seconds (e.g. one scan per minute or even one scan per hour) cannot be detected using derived "time based" features. To capture these types of the attacks, we also derived "connection based" features that capture the same characteristics of the connection records as time based features, but are computed for the last 100 connections.

It is well known that constructed features from the data content of the connections are more important when detecting R2L and U2R attack types, while "time-based' and "connection-based" features are more important for detection of DoS and probing attack types (Lee 1998).

6.2. Results for Learning from Rare Class

KDDCup'99 data set is an extension of DARPA'98 data set with a set of additionally constructed features. Unlike the data set that we have developed, it does not contain some basic information about the network connections (e.g. start time, IP addresses, ports, etc.) that we needed for our analysis of multi-connection attacks. The data set was mainly constructed for the purpose of applying data mining algorithms. Therefore, we have also used this data set as a testbed for our algorithms for learning from rare class. Two of five classes are considered rare, U2R and R2L classes respectively, with 0.4% and 5.7% of the entire population of data.

When experimenting with the SMOTEBoost algorithm, different values for the SMOTE parameter that controls the amount of generated examples, ranging between 100 and 500, were used for the minority classes. The values of SMOTE parameters for U2R class were higher than the SMOTE parameter values for R2L class, since R2L class is better represented in KDDCup 1999 data set than the U2R class (R2L has larger number of examples). Our experimental results also showed that the higher values of SMOTE parameters for R2L class could lead to overfitting and decreasing the prediction performance on that class. Fig. 4 shows the precision, recall and the F-value for the combination of SMOTE parameters that give the best classification performance of the SMOTEBoost algorithm.

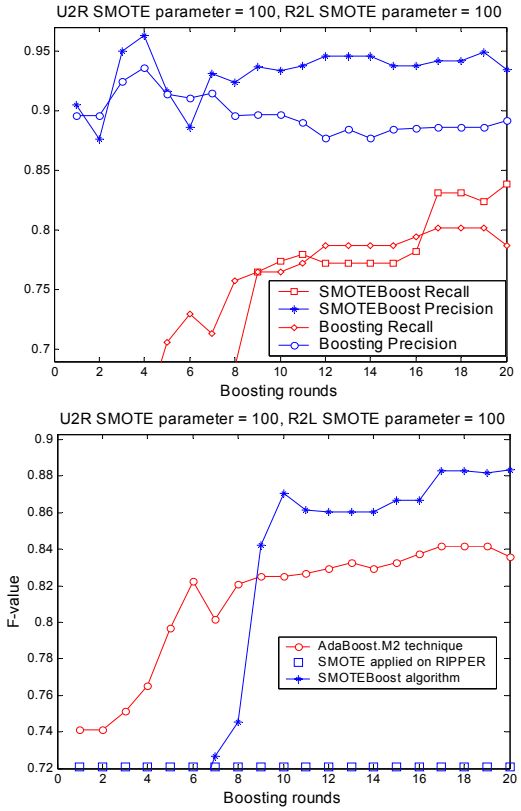

Fig. 4. *Precision*, *Recall*, and *F-values* for the U2R class

Table 2. Results of association-based classification algorithm on KDDCUP'99 data

Added features	Class	*Precision*	*Recall*	*F-value*
No added features	**U2R**	84.8%	57.4%	68.4%
	R2L	96.7%	75.9%	85.1%
High Precision	**U2R**	88.6%	68.4%	77.2%
	R2L	96.5%	78.9%	86.8%
High Recall	**U2R**	90.1%	73.5%	81.0%
	R2L	92.9%	75.9%	83.5%
High F-value	**U2R**	94.2%	83.1%	88.3%
	R2L	96.2%	84.3%	89.8%

When our proposed association based classification algorithm is applied on KDDCup data set, experimental results indicate that a ignificant increase in prediction performance may be achieved by considering not only the itemsets with high precision but also the itemsets with high recall and F-value. Table 2 shows the precision, recall and the F-value when no itemsets were added to the original data set as well as when the itemsets with high precision, recall and F-value were added as "meta-features" to the original data set.

6.3. Anomaly Detection Results on DARPA'98 Data

In order to perform our evaluation of both single-connection and multi-connection attacks, we applied presented anomaly detection algorithms to our data set constructed from DARPA'98 data. After the features are constructed and normalized, anomaly detection schemes were tested separately for the attack bursts, mixed bursty attacks and non-bursty attacks.

6.3.1. Evaluation of Bursty Attacks. Our experiments were first performed on the attack bursts. Fig. 5 illustrates the ROC curves of all proposed algorithms and show how the detection rate and false alarm rate vary when different thresholds are used. Since the *unsupervised SVM approach* was not able to achieve a false alarm rate of 1% and 2%, these results were omitted from the Fig. 5. Using the standard metrics, we consider a burst to be detected if the corresponding *burst detection rate* is greater than 50%. Since we have a total of 19 bursty attacks, overall detection rate in Fig. 5 was computed using this rule. It is apparent form Fig. 5 that the most consistent anomaly detection scheme is the *LOF approach*, since it is only slightly worse than the *NN approach* for low false alarm rates (1% and 2%), but significantly better than all other techniques for higher false alarm rates (greater than 2%). The *Mahalanobis-based approach* was consistently inferior to the *NN approach* and was able to detect only 11 multiple-connection attacks. This poor performance of *Mahalanobis-based scheme* was probably due to the fact that the normal behavior may have several types and cannot be characterized with a single distribution. In order to alleviate this problem, there is a need to partition the normal behavior into several more similar distributions and identify the anomalies according to the Mahalanobis distances to each of the distributions. However, there are also scenarios when these two schemes have different detecting behavior. Fig. 6 illustrates the detection of burst 2 from week 2 using *NN* and *LOF*. It is apparent that the *LOF approach* has a smaller number of connections that are above the threshold than the *NN* approach (smaller *burst detection rate*), but it also has a slightly better response performance than the *NN* approach for specified threshold. In addition, both schemes demonstrate some instability (low peaks) in the same regions of the attack bursts that are probably due to occasional "reset" value for the feature called "connection status". However, when detecting this bursty attack, the *NN approach* was superior to other two approaches. The dominance of the *NN approach* over the *LOF approach* probably lies in the fact that the connections of this type of attack (portsweep attack, probe category) are located in the sparse regions of the normal data, and the *LOF approach* is not able to detect them due to low density, while distances to their nearest neighbors are still rather high and therefore the *NN approach* was able to identify them as outliers. Finally, Fig. 6 evidently shows that in spite of the limitations of the *LOF approach* mentioned above, it was still able to detect the attack burst, but with higher instability which is penalized by larger surface area.

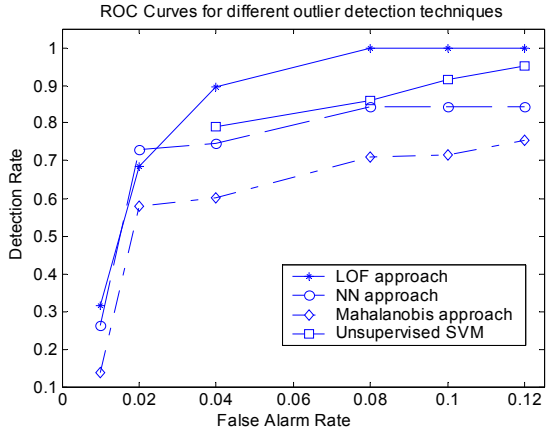

Fig. 5. ROC curves showing the performance of anomaly detection algorithms on bursty attacks

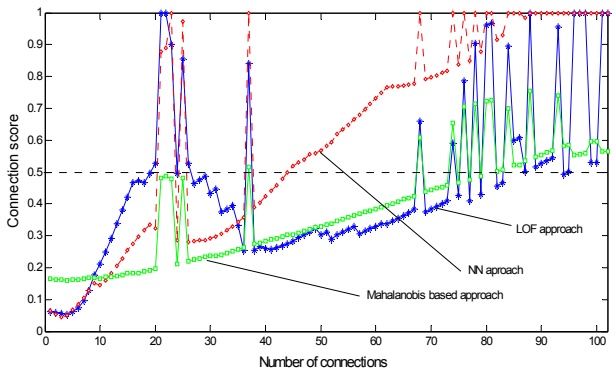

Fig. 6. The detection of bursty attack 2 from week 2

6.3.2. Evaluation of Single Connection Attacks. Fig. 7 shows the ROC curves of all the proposed anomaly detection schemes. The *LOF approach* was again superior to all other techniques and for all values of false alarm rate. All these results indicate that the *LOF* scheme may be more suitable than other schemes for detecting single connection attacks especially R2L intrusions, since for the fixed false alarm rate of 2%, the *LOF* approach was able to detect 7 out of 11 attacks, while the *NN* approach was able to pickup only one. Such superior performance of the *LOF approach* may be explained by the fact that majority of single connection attacks are located close to the dense regions of the normal data and thus not visible as outliers by *the NN approach*.

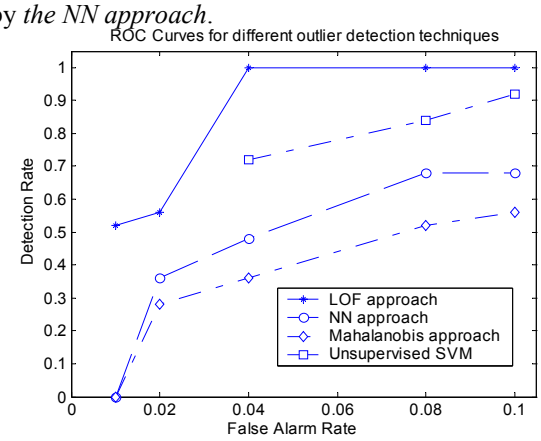

Fig. 7. ROC curves showing the performance of anomaly detection algorithms on single-connection attacks

6.4. Results from Real Network Data

Due to various limitations of DARPA'98 intrusion detection evaluation data discussed above [24], we have repeated our experiments on live network traffic at the University of Minnesota. When reporting results on real network data, we were not able to report the detection rate, false alarm rate and other evaluation metrics reported for DARPA'98 intrusion data, mainly due to difficulty to obtain the proper labeling of network connections.

Since we work on intrusion detection issues together with system administrators at the University of Minnesota, we could not apply all developed algorithms, but only the most prominent one. For this purpose we have selected the *LOF approach*, since it achieved the most successful results on publicly available DARPA'98 data set, especially in detecting single-connection attacks. The *LOF* technique showed also great promise in detecting novel intrusions in real networks and during the past few months it has been very successful in automatically identifying several novel intrusions at the University of Minnesota that could not be detected using state-of-the-art intrusion detection systems such as SNORT (SNORT IDS). Many of these attacks have been on the high-priority list of CERT/CC recently. Examples include:

- On August 9th, 2002, CERT/CC issued an alert "widespread scanning and possible denial of service activity targeted at the Microsoft-DS service on port 445/TCP" as a novel Denial of Service (DoS) attack that had not been observed before. In addition CERT/CC expressed "interest in receiving reports of this activity from sites with detailed logs and evidence of an attack." This type of attack was the top ranked outlier on August 13th, 2002, by our anomaly detection tool in its regular analysis of University of Minnesota traffic. This could not be detected by SNORT and other such tools since the port scanning was a low rate non-sequential one.
- On June 13th, 2002, CERT/CC sent an alert for an attack that was "scanning for an Oracle server". This can be a potentially insidious type of insider attack on databases. Our tool identified an instance of this attack on August 13th from the UM network flow data by listing it is as the second highest ranked outlier. This type of attack is difficult to detect using other techniques, since the Oracle scan is hidden within a high rate Web scan.
- On August 8th and 10th, 2002, our techniques identified machines running an illegal Microsoft PPTP VPN server, and an illegal FTP server, respectively – both as the top ranked outliers. The FTP attack did not have a known signature, and hence SNORT did not detect it. For the VPN attack, the collected GRE traffic is part of the normal traffic, and hence not analyzed by tools such as SNORT.
- On October 10th, 2002, our anomaly detection tool detected two activities of slaper worm which were not identified by SNORT since they were variations of existing warm code. These worms could be potentially identified by SNORT using possible rules, but the false alarm rate will be too high.
- On October 10th, 2002, distributed windows networking scan from two different source locations was identified by our technique. It is interesting to note that all the network connections associated with this attack were assigned the same anomaly score, which indicated that the connections belong to the same attack. Since this was also slow scanning activity, SNORT was not able to detect it. Using appropriate rules SNORT would be able to see two or three independent scanning attacks in the best case, but powerless to see a distributed attack.

CONCLUSION

It is crucial that we pay sufficient attention to making our information systems - especially those used for critical functions in the military and commercial sectors - resistant to and tolerant of such attacks. Research at the Army High Performance Computing Research Center is focusing on applying data mining to develop techniques that can be used to detect, and thwart from known as well as unknown cyber threats.

Our continuing objective is to develop an overall framework for defending against attacks and threats to computer systems. Data generated from network traffic monitoring tends to have very high volume, dimensionality and heterogeneity, making the performance of serial data mining algorithms unacceptable for on-line analysis. In addition, cyber attacks may be launched from several different locations and targeted to many different sources, thus creating a need to analyze network data from several networks in order to detect these distributed attacks. Therefore, development of new classification and anomaly detection algorithms that can take advantage of high performance computers and be computationally tractable for on-line and distributed intrusion detection is a key component of this project. To detect known attacks, our approach will use the public-domain signature-based techniques, while unknown and novel attacks will be detected using our anomaly detection schemes. In addition, the system will have a visualization tool to aid the analyst in better understanding anomalous/suspicious behavior detected using our techniques. We believe this tool will significantly enhance the capability of analysts responsible for cyber threat prevention. In addition, we plan to extend our research in applying data mining for other security aspects including prevention from cyber attacks, recovery from them, identifying new system vulnerabilities and setting new policy mechanisms.

ACKNOWLEDGMENTS

The authors are grateful to Richard Lippmann and Daniel Barbara for providing data sets. This work was supported by Army High Performance Computing Research Center contract number DAAD19-01-2-0014. The content of the work does not necessarily reflect the position or policy of the government and no official endorsement should be inferred. Access to computing facilities was provided by the AHPCRC and the Minnesota Supercomputing Institute.

REFERENCES

D. Barbara, N. Wu, S. Jajodia, Detecting Novel Network Intrusions Using Bayes Estimators, *First SIAM Conference on Data Mining*, Chicago, IL, 2001.

M. M. Breunig, H.-P. Kriegel, R. T. Ng, J. Sander, LOF: Identifying Density-Based Local Outliers, *Proceedings of the ACM SIGMOD Conference*, 2000.

D. Denning, An Intrusion Detection Model, *IEEE Transactions on Software Engineering*, SE-13:222-232, 1987.

M. Joshi, V. Kumar, R. Agarwal, Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements, *First IEEE International Conference on Data Mining*, San Jose, CA, 2001.

M. Joshi, R. Agarwal, V. Kumar, Predicting Rare Classes: Can Boosting Make Any Weak Learner Strong?, *Proceedings of Eight ACM Conference ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Canada, 2002.

M. Joshi, R. Agarwal, V. Kumar, PNrule, Mining Needles in a Haystack: Classifying Rare Classes via Two-Phase Rule Induction, *Proceedings of ACM SIGMOD Conference on Management of Data*, May 2001.

M. Joshi, V. Kumar, CREDOS: Classification using Ripple Down Structure (A Case for Rare Classes), in review.

A. Lazarevic, N. Chawla, L. Hall, K. Bowyer, SMOTE-Boost: Improving the Prediction of Minority Class in Boosting, *AHPCRC Technical Report*, 2002.

A. Lazarevic, A. Ozgur, L. Ertoz, J. Srivastava, V. Kumar, A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection, in review.

W. Lee, S. J. Stolfo, Data Mining Approaches for Intrusion Detection, *Proceedings of the 1998 USENIX Security Symposium*, 1998.

R. P. Lippmann, D. Fried, I. Graf, J. W. Haines, K. P. Kendall, D. McClung, D. Weber, S. E. Webster, D. Wyschogrod, R. K. Cunningham, and M. A. Zissman, Evaluating Intrusion Detection Systems: The 1998 DARPA Off-line Intrusion Detection Evaluation, *Proceedings DARPA Information Survivability Conference and Exposition (DISCEX)*, Vol 2, 12-26, Los Alamitos, CA, 2000.

P.C. Mahalanobis, On Tests and Meassures of Groups Divergence, *International Journal of the Asiatic Society of Benagal*, 26:541, 1930.

J. McHugh, The 1998 Lincoln Laboratory IDS Evaluation (A Critique), *Proceedings of the Recent Advances in Intrusion Detection*, 145-161, Toulouse, France, 2000.

B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, R. Williamson, Estimating the Support of a High-dimensional Distribution, *Neural Computation*, vol. 13, no. 7, 1443 1471, 2001.

SNORT Intrusion Detection System. www.snort.org.

Successful Real-Time Security Monitoring, Riptech Inc. white paper, September 2001.

Tcptrace software tool, www.tcptrace.org.

United States White Paper: Concepts for the Objective Force, 2001.

F. Vizard, Waging War.com: A Hacker Attack Against NATO Spawns a War in Cyberspace, Popular Science, p. 80, July 1999.