International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Review on Intrusion Detection Using Fuzzy ARTMAP with Feature Selection Technique

Swati Sonawale[1], Roshani Ade[2]

[1]Department of computer Engineering, Dr.D.Y.Patil School of Engineering & Technology, Savitribai Phule Pune University, Pune, India

[2]Assistant Professor, Department of computer Engineering, Dr.D.Y.Patil School of Engineering & Technology
Savitribai Phule Pune University, Pune, India

Abstract: ***Considerable research work have been conducted towards Intrusion Detection Systems (IDSs) as well as feature selection. IDS guard a system from attack, misuse, and compromise. It can also screen network activity. Network traffic observing and extent is increasingly regarded as an vital role for understanding and improving the performance and security of our cyber infrastructure. In this research we have proposed framework by using advance feature selection technique & by using dimensionality reduction technique we can reduce IDS data then applying Fuzzy ARTMAP classifier we can find intrusions so that we get accurate results within less time. This technique is very efficient as it saves time as well as storage space.***

Keywords: Feature Selection, Intrusion, Redundancy, Fuzzy ARTMAP.

1. Introduction

For any business security is very vital because it may encompass confidential information, sensitive data. Reputation of business depends on it because if any security breaks happen it can damage millions of data within few seconds. So there is necessity of intrusion detection system so that if any intrusion happens we can take curative actions. It is a key method which will help to find various types of attacks.in a network system. There are following types of intrusion system.

IDSs are divided into two broad groups: host-based (HIDS) and network-based (NIDS). A host-based IDS needs small programs (or agents) to be installed on every systems to be administered. The agents screen the operating system and write down records to log records and/or activate alarms. A network-based Intrusion Detection System typically includes of a network application with a Network Interface Card (NIC) functioning in promiscuous mode and a separate management of interface. IDS is positioned on a network segment or boundary and monitor all traffic on that division. The present trend in intrusion detection is to combine both host based and network based information to develop hybrid systems that have more efficient.[1]

1.1 Host Based Intrusion Detection (HIDS)

Host based intrusion detection (HIDS) states to intrusion detection that takes place on a single host system. The data is collected from an single host system. The HIDS agent screens activities such as integrity of system, application action, file changes, host based network traffic, and system logs.

By using common hashing tools, file timestamps, system logs, and monitors system calls and the local network interface gives the agent insight to the present state of the local host. If there is any illegal change or activity is observed, it alerts the user by a pop-up, it alerts the central management server, blocks the activity, or a combination of the above three. The decision should be based on the policy that is installed on the local system. These host-based procedures are measured the passive component.[1]

1.2 Network Based Intrusion Detection (NIDS)

A network-based intrusion detection system (NIDS) is used to screen and examines network traffic to guard a system from network-based threats where the data is traffic across the network. A NIDS tries to detect cruel activities such as denial-of-service (Dos) attacks, port scans and observing the network traffic attacks. NIDS contains a number of sensors to screen packet traffic, one or more than servers for NIDS management functions, and one or more management relieves for the human interface. NIDS inspects the traffic packet by packet in real time, or near to real time, for trying to discover intrusion patterns. The analysis of traffic patterns to identify intrusions may be done at the sensors, at the administration servers, or mixture of the both. These network-based measures are considered the active element. [2]

2. IDS Classification Technique

There are various algorithms available for classification of Intrusion detection system Such as follows:

2.1 Support Vector Machine

Support Vector Machines are built on the idea of decision planes that describe decision boundaries. A decision plane is one that divides a set of objects having different class relationships. A graphic illustration is shown in below figure 2 [3] In this model, the objects belong either to class GREEN or RED. The splitting line states a boundary on the right side of which all objects are GREEN and to the left of which all objects are RED. Any new object falling to the right is labeled, i.e., categorized, as GREEN (or classified as RED should it fall to the left of the splitting line).

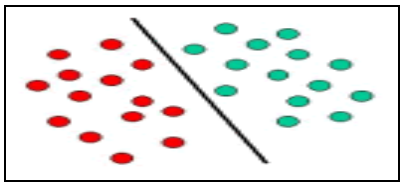

Figure 2: Partition of objects based on different classes

The above is a classic instance of a linear classifier, i.e. a classifier that splits a set of objects into their relevant groups (GREEN and RED in this case) with a line. Most cataloging tasks, however, are not that simple, and frequently more difficult assemblies are necessary in order to make an ideal separation.

2.2 Decision Trees

It has three basic elements:
1. A decision node
2. A leaf node which contains name of the class to which it belongs.
3. An edge that may be output of test

Following points are important for decision tree. By using given training set we have to build decision tree We have to select test attribute for each node as well as labelings for Sorting will start from root node. By using this node we will test the attribute. This will allows to move downward of tree.

1) This process will be continuous till the leaf node. Thus instances are classified.
2) The decision trees select the best features for each decision node during the construction of the tree

2.3 Arificial Neural Network

It is also called as a neural network. It consist of large number of neurons functioning together to solve specific problem. Output of each neuron is given as a input to the next layer. Artificial neural network is an information processing model that is encouraged by biological nervous system .It is composed of a large number of highly interconnected processing elements called neurons. An artificial neural network is constituted for specific application such as pattern recognition or data classification. Supervised learning is used when we have a set of training data. This training data encompasses of some input data that is linked with some accurate output values. The output values are frequently stated as a target values. This training data is used by learning algorithms like back propagation or genetic algorithms Back propagation uses the target values to calculate the mean square error of the artificial neural network and genetic algorithms practice target values when calculating the fitness levels of an individual in a population.

But the aim of the learning algorithm is not to produce a neural network that outputs perfect values for the training data the mission is to give good values for input data that is from the real world and not from the training set. When we train the network to hard against the training set we tend to learn the 'noise' in the measured data and not the underlying structure, we don't realize the 'whole picture', this is called over fitting. If we divided our training data into two pieces and uses one for training and the other part for validation we will see that both the error for the training data and the error of the validation data will decrease at first, then at some point we starts to over fit the network and we see that the training data still gets better values but the validation error get bigger again, this is the point of over fitting and this is when we should stop the training.

Applications of Neural Network

1) It has ability to derive meaning from complicated or imprecise data.
2) It removes patterns & identifies trends that are too challenging to be notified by either humans or other computer techniques.
3) Adaptive learning
4) Real time operation

3. Feature Selection

It is a method that Select the most "significant" subset of attributes according to some selection criteria. It means select minimum subset of m features from the original set of n features so that feature space optimally decreases conferring to definite assessment criteria.

Feature selection has been active research area in pattern recognition statistics and data mining groups. The main concept of feature selection is to select subset of input variable by eliminating feature with no predictive information .e.g. Physician may make decision based on particular feature whether a dangerous operation is necessary for treatment or not . Feature selection can be done two ways.

A. Supervised learning- in this higher classification accurateness is expected.
B. Unsupervised learning- is designed to find natural grouping to form high quality of cluster. It encompasses high dimensional data which contains redundant information.[4] Due to this following consequences may occur:
   1) It reduces the accuracy of data mining algorithms
   2) It can slow down the mining process
   3) There may be a problem in storage and retrieval
   4) It is hard to interpret

4. Problem Statement

The task of feature selection becomes more challenging with the small-labeled-sample problem [5] in which the extent of data that is unlabeled can be much greater than the amount of labeled data On the other hand, supervised feature selection algorithms need a big extent of labeled training data. As a result, such algorithms provide inadequate information about the structure of the target model, and can thus fail to detect the related features that are discriminative to different classes. On the other hand, unsupervised feature selection algorithms reject label information and thus may lead to performance worsening

We propose a system to offer dimensionality reduction with semi-supervised feature selection with constraint, relevance, and redundancy.

5. Architecture of Feature Selection with IDS

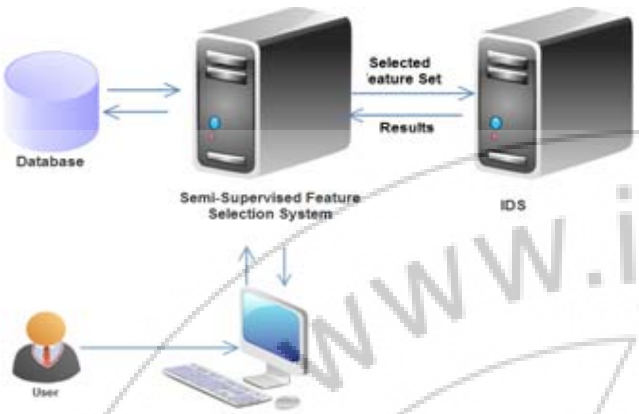

Figure 1: Architecture of feature selection with intrusion detection system

Architecture of our system is as shown in above figure. Here we are taking kdd99 dataset as a input & then applying semi-supervised feature selection algorithm we are choosing only selected features. By using that features we can find whether there is attack in the system or not. Thus we will reduce data & by using reduction in data we are saving processing time & we get more accurate results within less time.

6. Fuzzy ARTMAP

Among various types of ART networks, Fuzzy ARTMAP (FAM) (Carpenter et al, 1992) [6] has appeared as a powerful supervised ART-based model for solving feature selection problems (Obaidat and Saudon, 1997; Lee and Tsai, 1998; Heinke and Hamker, 1998; Aggarwal et al, 1999). FAM combines the salient properties of ART with fuzzy set theory. FAM is very fast in training (Carpenter and Grossberg, 1994; Carpenter et al, 1995). As related with a large number of training times required in another NN model(e.g. MLP),FAM needs moderately few training time, which can be conducted incrementally.FAM also is proven to be noise tolerant (Charalampidis and Kasparis, 2001). In the context of pattern classification, FAM has been shown to produce good performance in a number of benchmark classification tasks (Carpenter and Grossberg, 1994, 1995). FAM has also been useful to tackle various real-world applications involving pattern classification with good performance, such as medical diagnostic , fault detection (Aggarwal et al, 1999; De & Chatterjee, 2004; Tan and Lim, 2004),manufacturing decision support system (Tan et al, 2005) and biometrics (Obaidat and Saudon, 1997; Lim and Woo, 2006).The above features make FAM an attractive NN model for investigation into the problem of feature selection.

7. KDD99 as a Dataset

For the purpose of this study we have used kdd99 as a dataset. The main goal is to use this dataset is that this database is freely available.

Since 1999, KDD'99 has been the most wildly used data set for the estimation of anomaly detection techniques. This data set is prepared by Stolfo et al. and is made based on the data captured in DARPA'98 IDS estimation Program. DARPA'98 is around 4 gigabytes of compacted raw (binary) tcp dump data of 7 weeks of network traffic, which can be managed into about 5 million connection records, each with around 100 bytes. The two weeks of test data have around 2 million link records. KDD training dataset comprises of nearly 4,900,000 particular connection vectors each of which contains 41 features and is labeled as either normal or an attack, with just one specific attack type. The simulated attack comes under one of the following four attacks:

Table1: Basic characteristics of the KDD 99 intrusion detection datasets in terms of samples

Dataset	*DoS*	*Probe*	*u2r*	*r2l*	*Normal*
"10% KDD"	391458	4107	52	1126	97277
"Whole KDD"	3883370	41102	52	1126	972780

1) **Denial of Service Attack (DoS):** is an attack in which the attacker creates some computing or memory resource too busy or too full to handle valid requests, or denies valid users access to a machine.
2) **User to Root Attack (U2R)**: is a class of action in which the attacker starts out with entry to a normal user account on the system (perhaps gained by sniffing passwords, a dictionary attack, or social engineering and is capable to exploit some susceptibility to gain root access to the system.
3) **Remote to Local Attack (R2L)**: happens when an attacker who has the ability to send packets to a machine over a network but who does not have an account on that machine exploits specific weakness to gain local entrée as a user of that machine.
4) **Probing Attack**: is an task to gather information about a network of computers for the apparent resolution of circumventing its security controls.[21]

8. Conclusion and Future Enhancement

Security tools installation, monitoring to ensure security is the responsibility of the Security Administrator in an organization. IDS generate a large number of alerts (false positives). Most of these alerts demand manual intervention from Administrator. Continuous monitoring of alerts and there by evolving judgment for improving security is the major concern. Thus, Feature relevance analysis is performed on KDD 99 training set, which is widely used by machine learning researchers. Feature relevance is expressed in terms of information gain, which gets higher as the feature gets more discriminative. Thus by using this technique of feature selection & fuzzy art map we get accurate results within less time & it gives efficient solution. In future we can use another classifier for classification of intrusion detection system so that it will get more accurate results & it will reduce time.

References

[1] Intrusion Detection Technique by Using Fuzzy ART on Computer Network Security 978-1-4577-2119-9/12 2011 IEEE

[2] Performance Evaluation of the Fuzzy ARTMAP for Network Intrusion Detection S.M. Thampi et al. (Eds.): SNDS 2012, CCIS 335, pp. 23–34, 2012.
[3] Z. Zhao and H. Liu, Spectral Feature Selection for Data Mining (Data Mining & Knowledge Discovery Series). Boca Raton, FL, USA:Chapman and Hall-CRC, 2012.
[4] Efficient Semi-Supervised Feature Selection: Constraint, Relevance, and Redundancy Khalid Benabdeslem and Mohammed Hindawi VOL. 26, NO. 5, MAY 2014
[5] L.Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy" J. Mach. Learn. Res., vol. 5, pp. 1205 1224,Oct. 2004.
[6] Performance Evaluation of the Fuzzy ARTMAP for Network Intrusion Detection S.M. Thampi et al.(Eds.):SNDS2012,CCI335,pp. 23–34, 2012.
[7] Z. Zhao and H. Liu, "Semi-supervised feature selection via spectral analysis," in Proc. SIAM Int. Conf. Data Mining, Tempe,AZ, USA, 2 007, pp. 641–646.
[8] M. Kalakech, P. Biela, L. Macaire, and D. Hamad, "Constraint scores for semi-supervised feature selection: A comparative study," Pattern Recognition. Lett., vol. 32, no. 5, pp. 656–665, 2011.
[9] K. Benabdeslem and M. Hindawi, "Constrained Laplacian score for semi-supervised feature selection," in Proc. ECML-PKDD, Athens, Greece, 2011, pp. 204–218.
[10] K. Allab and K. Benabdeslem, "Constraint selection for semisupervised topological clustering," in Proc. ECML-PKDD, Athens, Greece, 2011, pp. 28–43.
[11] C. Ding and H. C. Peng, "Minimum redundancy feature selection from microarray gene expression data," in Proc. IEEE CSB, 2003, pp. 523–528.
[12] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
[13] B. Auffarth, M. Lopez, and J. Cerquides, "Comparison of redundancy and relevance measures for feature selection in tissue classification of CT images," in Proc. 10th ICDM,Berlin, Germany,2010, pp. 248–262.
[14] M. Hindawi, K. Allab K. Benabdeslem, "Constraint selection based semi-supervised feature selection," in Proc. IEEE ICDM,Vancouver, BC, Canada, 2011, pp. 1080–1085
[15] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. 5th Symp. Math. Statist. Probab., Berkley USA, 1967, pp. 281–297.
[16] J. H. Ward, "Hierarchical grouping to optimize an objective function," J. Amer. Statist. Assoc., vol. 58, no. 301, pp. 236–244,1963.
[17] M. Kalyani and M. Sushmita, "Clustering and its validation in a symbolic framework," Pattern Recognit. Lett., vol. 24, no. 14,pp. 2367–2376, 2003.
[18] X. He, D. Cai, S. Yan, and H. Jiang Zhang, "Neighborhood preserving," in Proc.10th IEEE Int. Conf. Comput. Vision,Beijing, Germany, 2005, pp. 1208–1213.
[19] I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.