Network Intrusion Detection Using Supervised Machine Learning Technique

Nivedita S Naganhalli, Dr Sujata Terdal

Abstract— In the modern computer world, use of the internet is increasing day by day. However, the increasing use of the internet creates some security issues. These days, such new type of security attacks occurs everyday and it is not easy to detect and prevent those attacks effectively. One common method of attack involves sending large amount of request to site or server and server will be unable to handle such huge requests and site will be offline for many days . This type of attack is called distributed denial of service (DDOS) attack, which act as a major security threat to internet services and most critical attack for cyber security world . Detection and prevention of Distributed Denial of Service Attack (DDoS) becomes a crucial process for the commercial organizations that uses the internet . Different approaches have been adopted to process traffic information collected by a monitoring stations (Routers and Servers) to distinguish malicious traffic such as DDoS attack from normal traffic in Intrusion Detection Systems (IDS). In general, Machine learning techniques can be designed and implemented with the intrusion systems to protect the organizations from malicious traffic. Specifically, supervised clustering techniques allow to effectively distinguish the normal traffic from malicious traffic with good accuracy. In this paper, machine learning algorithms are used to detect DDoS attacks collected from "KDDcup 99 Dataset" , pre-processing and feature selection technique is used on the dataset to enhance the performance of the classifiers and reduce the detection time. The classification algorithms such as C4.5 decision tree and NavieBayes is applied on the training dataset and the implementation of the algorithm is done using spyder tool. The performance comparison of algorithms is shown using confusion matrix and it is found that C4.5 decision is more efficient in detection of DDOS attack .The proposed method can be used as DDoS defense system.

Key Words: C 4.5Decision Tree, DoS attack detection, IDS, KDD Dataset,Naive Bayesian classifier ,Machine learning.

1 INTRODUCTION

With the rapid development of information technology , the Computer networks are widely used by industry, business and various fields of the human life[1]. As a result, it is very important for IT administrators to create a trusted network. In addition, the rapid development of information technology has created some problems in building a reliable network. This is a very difficult task. Many types of attacks are threatening the availability, integrity, and privacy of computer networks. A Denial of Service (DOS) attack is considered one of the most common attacks. The purpose of a DOS attack is to temporarily deny multiple end-user services. Typically, network resources are typically consumed and unwanted request systems are overloaded. For this reason DOS acts as a large umbrella for all types of attacks which aim to consume computer and network resources. Hence, it is very difficult to detect all types of attacks hence the intrusion detection system (IDS) has became an essential part of network security. It is implemented to monitor network traffic in order to generate alerts when any attacks appear. IDS can be implemented to monitor network traffic of a specific device (host intrusion detection system) or to monitor all network traffics (network intrusion detection system) which is the common type used. In general, there are two types of IDS (anomaly base or misuse base).[2] Anomaly based intrusion detection system is implemented to detect attacks based on recorded normal behavior. As a result, we compare current live traffic with previously recorded normal traffic. This type of intrusion detection system is widely used because it can detect new types of intrusions. But from another perspective, record the most important values of a false positive alarm. This means that there are a number of legitimate packages considered to be attack packages. However, an exploit intrusion detection system is implemented to detect attacks based on attack signature stores. There are no false alarms, but at the same time new types of attacks (new signatures) can successfully send alerts. As the existing intrusion detection systems require input from human which is expensive to determine the target is either normal or attack packet ,machine learning algorithms[2] can be used as an alternative to discover appropriate behavior as normal and attack. Recently, there has been an increased interest in machine learning approaches to build intrusion detection models. Using machine learning approaches for IDS in which intrusion detection is considered as a classification problem, that is identifying normal and other types of intrusive behavior. An accurate intrusion detection model can be built by choosing an effective classification approach. Hence in this paper, the proposed method uses machine learning classifiers such as C4.5 Decision Tree and Naive Bayes classifiers to evaluate and accurate the model of intrusion detection system based on a Knowledge Discovery in Databases (KDD) dataset. This paper is organized as follows : Section 2 presents the related overview of machine learning techniques for IDS . whereas proposed method is introduced in Section 3. Then, the experimental results are discussed in Section 4. Finally, conclusions is mentioned in Section 5.

2 RELATED WORK

A security mechanism used to monitor the abnormal behavior of the network is an Intrusion Detection System(IDS). The IDS identifies and informs that whether the user activity is normal or not. The users activities are [3]compared by the IDS with the already stored intrusion records to identify the intrusion. IDS learns the patterns by the training data, so the misuse based method is used. Misuse based detection can detect only the known attack, new attacks cannot be identified. Anomaly based IDS observes the normal behavior and if there is a change in the behavior then it considers that behavior as anomaly. So anomaly based IDS can detect new attacks that are not learned from the training model. Accurate predictive models can be built for large data sets using supervised machine learning techniques. In supervised learning ,learning data comes with labels or desired outputs and the objective is to find a general rule that maps inputs to outputs. This kind of learning data is called labeled data .The learned Rule is then used to label new data with unknown outputs. It involves building a machine learning model based that is based on labeled samples.

There are several supervised learning algorithms. Artificial Neural Network, Bayesian Statistics, Gaussian Process Regression, Lazy learning, Nearest Neighbor algorithm, Support Vector Machine, Hidden Markov Model, Bayesian Networks, Decision Trees(C4.5,ID3, CART, Random Forrest), K-nearest neighbor, Boosting, Ensembles classifiers (Bagging, Boosting), Linear Classifiers (Logistic regression, Fisher Linear discriminant, Naive Bayes classifier, Perceptron, SVM), Quadratic classifiers are some of the most popular supervised learning algorithms

A. Naïve Bayes

Naive Bayes is based on the Bayesian method for performing the classification process. It is a simple and easiest technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. The new proposal was innovative as Hidden Naïve bayes which shows more advantage then traditional naïve bayes [4] The proposed method in the paper use of Hidden Naïve Bayes (HNB) provides more accurate results than the traditional Naïve Bayes model. Hidden Naive Bayes (HNB) model can be applied to intrusion detection problems (DOS attacks) that suffer from dimensionality highly correlated features and high network Data stream volumes [4]. In a data mining model that loosens the naive Bayes methods Conditional impartiality assumption. The proposed work addresses the attack detection in network layer [5] in this paper a new dataset was collected that consist of DDOS attacks in different network layers. DDoS attacks are detected using three techniques Multilayer perceptron (MLP), Naïve Bayes and Random Forest. MLP showed the highest accuracy rate (98.63%) as compared to other techniques. [6] An hierarchical layered approach was proposed to increase the detection rate of attacks. Model used Naive Bayes classifier with K2 learning process on reduced NSL KDD dataset for each attack class. In the proposed model every layer is trained to detect a single type of attack. The outcome of one layer is passed on to another layer to increase the detection rate.a new hybrid algorithm[7] for adaptive network intrusion detection using Naive Bayesian classifier and ID3 algorithm, which analyzes the large volume of network data and considers the complex properties of attack behaviours to improve the performance of detection speed and detection accuracy. In[8] it is found that Naïve Bayes (NB) can perform very well when moderate dependencies exist in the data. It has been shown that the performance of Naïve Bayes classifier improves when redundant features are removed.The [9] paper encompasses incorporate flow correlation analysis along with Naïve Bayesian classification process in order to determine the intruded packets in the network. Since the classification scheme is based on posterior conditional probabilities, it identifies attacks that occur in an uncertain situation .The results show that the proposed scheme can effectively classify packets than existing classification models.

B. Decision Trees

Decision tree is one of the simple technique used in the machine learning and data mining. It is used as a predictive model in which observations about an item are mapped to conclusions about the item's target value. In the process of decision analysis, a decision tree can be used to represent decisions and decision making visually and explicitly. In this algorithm, the data set is learnt and modelled. Therefore, whenever a new data item is given for classification, it will be classified accordingly learned from the previous dataset . Decision Tree algorithm can also be used for DOS attack detection. [10]The data mining approach to detect DOS attacks, using classification techniques. The above approach is used to classifying “normal” traffic against “abnormal” traffic in the sense of DoS attacks. The paper evaluates the performance of J48 decision tree algorithm for the detection of DoS attacks and then compares it with two rule based algorithms which are OneR and Decision table. In[11] designed a DDoS-detection system based on a decision-tree technique in which after an attack is detected, the system trace back to the attacker’s locations using a traffic-flow pattern-matching technique. A C4.5 classifier is used for detection of dos attacks. [11] in this paper proposed a learning algorithm for anomaly based network intrusion detection system that distinguishes attacks from normal behaviors and identifies different types of intrusions using decision tree algorithm. Data set used is KDD99 benchmark network intrusion detection dataset. The classes in KDD99 dataset categorized into one normal class and four intrusion classes: probe, DOS, U2R, and R2L.In[12] proposed decision tree to reduce the probability of over fitting the training data. Decision trees (DTs) are popular in misuse detection systems, as they yield good performance and offer some benefits over other machine learning techniques.[13] have examined the performance of several machine learning techniques including C4.5 DT. The DT obtained good accuracy, but does not perform as well as other techniques on some classes of intrusion, particularly U2R and R2L attacks, both of which are minor classes and include a large proportion of new attack types. Similar observations have been made by [14] and demonstrated that DTs are very sensitive to the training data and do not learn

well from imbalanced data. Furthermore, they found that DTs and Random Forests (ensemble of DTs) are very sensitive to the data selected for training, i.e., the performance varied significantly on different
folds (subsets) of the data. In[15] also proposed a modification to the C4.5 DT classifier, aimed at reducing the false positive rate. They changed the way in which the trees are built, by taking into account the type of errors that may be produced, choosing attributes that are less likely to produce false positives. The modified C4.5 DT outperformed the original DT and the sampling approach.

3 PROPOSED METHOD

The workflow of the proposed method is setup as shown in the Figure 1, starting with data collection (KDD-99 Dataset), Pre-Processing: Training and testing dataset, building model and result analysis

3.1 KDD cup 1999 dataset Collection

In 1998, the DARPA Intrusion Detection Assessment Program was prepared and managed by MIT Lincoln Labs. Its purpose was to study and evaluate intrusion detection research. Standard data sets include various simulation intrusions in military network environments. The connection to the dataset includes a sequence of TCP packets beginning and ending at a well-defined time between the source IP address and the destination IP address using a well-defined protocol. Each connection is categorized as a normal or specific type of attack. Data sets are categorized into five sub-sets: denial-of-service attacks, local or remote network attacks, user / root attacks, sample attacks, and generic data. Each record is classified as normal or attack with exactly one type of attack. They are categorized as follows:

- Denial of service (DoS) Denial of Service (DOS) allows a legitimate user to gain access to the machine by creating too much or too much computer resources or memory for an attacker to handle legitimate requests.
- R2L (Local Remote Attack (User)) Local Remote Attack (R2L) is a type of attack in which an attacker can send packets to a computer over the network and then exploit a vulnerability in the computer to illegally attack local access. On the machine.
- Root User Attack (U2R) Root User Attack (U2R) is the attack class that an attacker first accesses a regular user account on a system. The vulnerability could be exploited to gain root access to the system.
- Monitoring (monitoring and other discovery) detection is an attack type in which an attacker scans the network for known information or vulnerabilities. An attacker with a map of systems and services available on the network will use the information to detect attacks.

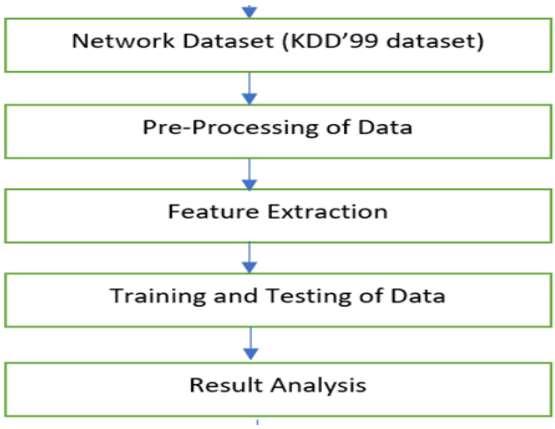

Figure 1 : System design

3.2 Data Pre-processing :

Pre processing involves
Handling Null Values :- isnull() method to check whether a null values is present in dataset .
Label Encoder: le=label Encoder() method label_encoder is used to transferring Categorical data into Numerical data

3.3 Feature Selection

Information Gain Ratio based feature selection: Features selected based on only information gain is biased towards attributes having many values. Information Gain Ratio (IGR) based Feature Selection removes this drawback by taking the splitting information of an attribute into account. Splitting information of an attribute is the entropy of pattern distribution into branches. Gain ratio of attribute decreases as value of split information increases.
Algorithm:

1. Start with the full set of attributes (set containing all attributes of the dataset) and null selected feature set.
2. Calculate information gain ratio of each attribute.
3. Choose an attribute from the total set with the highest information gain ratio.
4. Split the dataset into sub datasets depending on the attribute values.
5. Add the attribute to selected feature set and remove from set of attributes.
6. Repeat step 2 to 5 for each of the sub-datasets with the set of attributes, if instance in a sub-dataset belongs to more than one class.
7. Output the selected feature set.

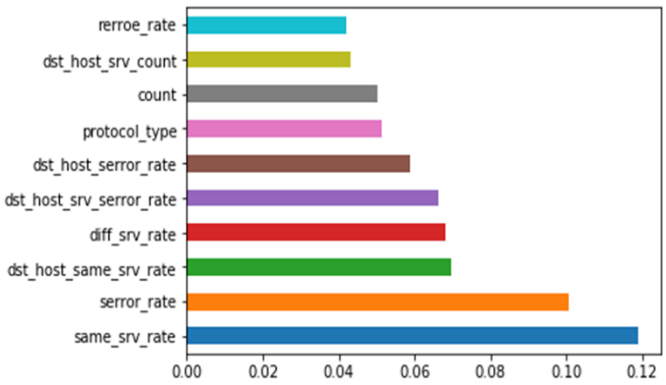

Figure 2 : Feature Selection

3.4 Model building

The supervised machine learning algorithms[1] are those algorithms which needs external assistance. The input dataset is divided into train and test dataset. The train dataset has output variable which needs to be predicted or classified. All algorithms learn some kind of patterns from the training dataset and apply them to the test dataset for prediction or classification. The proposed method distinguish the normal traffic from malicious traffic using supervised algorithms such as C4.5 Decision tree and Naïve Bayesian Classifier

a. C4.5 Decision tree

C4.5 Decision tree is one of the simple technique used in the machine learning and data mining. It is used as a predictive model in which observations about an item are mapped to conclusions about the item's target value. In the process of decision analysis, a decision tree can be used to represent decisions and decision making visually and explicitly.In this algorithm, the data set is learnt and modelled. Therefore, whenever a new data item is given for classification, it will be classified accordingly learned from the previous dataset

The steps of the algorithm are as follows:

1. If all the given training examples belong to the same class, then a leaf node is created for the decision tree by choosing that class.
2. For every feature 'a', calculate the gain ratio by dividing the information gain of an attribute with splitting value of that attribute. The formula for gain ratio is GainRatioa(a) = IG(a) / Split(a) where, S is the set of all the examples in the given training set.
3. Information gain of an attribute is computed as

$$IG(a) = Ent(S) - \sum_{a\_val \in values(a)} \frac{|S\_a|}{|a|} * Ent(S\_a)$$

where, S_a is the subset of S, values (a) is the set of all possible values of attribute 'a' and |a| is the total number of values in attribute 'a'

4. Entropy can be calculated as

$$Ent(S) = -\sum_{J=1}^{Num\_class} \frac{freq(Lj.S)}{|s|} * \log2 \frac{(freq(Lj,S)}{|S|}$$

where, L = L1,L2,…, Ln is the set of classes, and num_class is the number of distinct classes.

For our consideration num_class has only two values, namely, 'normal', and 'anomaly'.

5. Split value of an attribute is chosen by taking the average of all the values in the domain at that particular attribute. It can be formulated as

$$Split(a) = \left(\sum_{i=1}^{m} (a\_val)\, i \right)$$

where m is the number of values of an attribute 'a'.

6. Find the attribute with the highest gain ratio. Suppose, the highest gain ratio is for the attribute 'a_best'.
7. Construct a decision node that divides the dataset on the attribute 'a_best'.
8. Repeat steps from 1 to 4 on each subsets produced by dividing the set on attribute 'a_best' and insert those nodes as descendant of parent node.

C4.5 algorithm uses the following function for calculating the split value of an attribute

$$Split(a) = - \sum_{a\_val \in values(a)} \frac{|S - a|}{|a|} * \log2 \left(\frac{|S - a|}{|a|} \right)$$

b. Naive Bayesian Classifier

Naïve Bayesian classifier is a simple classification scheme, which estimates the class-conditional probability by assuming that the attributes are conditionally independent, given the class label c. The conditional independence assumption can be formally stated as follows:

$$P(A \setminus C = c) = \prod P(Ai \setminus C = c)$$

Where each attribute set A = {A1,A2,….,An} consists of n attribute The steps of the algorithm are as follows:

1.Input dataset Di , where n is the total number of training examples.

2.Calculate the prior probability P(Cj) for each class Cj in Dataset Di :

$$P(Cj) = \frac{\sum ti \rightarrow cj}{\sum_{i=1}^{n} ti}$$

3.Calculate the class conditional probabitities P(Aij\Cj) for each attribute values in dataset Di .

$$P(Aij \setminus Cj) = \frac{\sum_{i=1}^{n} Ai \rightarrow Cj}{\sum ti \rightarrow Cj}$$

4. Classify each training example ti in training data D with maximum posterior probabilities.

$$P(ei \mid Cj) = p(Cj) \prod_{k=1} \rightarrow p\, P(Aij \mid Cj)$$

5. Repeat steps 2 to 4 untill all the training examples ti in D are correctly classified.

4. DISCUSSION AND RESULTS

After the creation of the training models, the next step is the testing phase process implementation. There are several evaluations metrics can be used in a classification algorithm. In this paper, the confusion matrixes were generated for each machine learning classifiers. It includes significant information about existing and predicted output classes.It is the summary of prediction results on a classification problem. Confusion matrix is shown in Table which is the basis for checking the accuracy and Parameter of the proposed model.

Parameter	C4.5 decision model	Naïvebayes model
Accuracy	93.75 %	75%
Kappa statistic	0.875	0.500
Mean Absolute error	0.0625	0.2500
Root mean square error	0.0625	0.2500

Table : 1 *Camparsioon chart between c4.5 decision and Naivebayes*

The performance evaluation of the experiment is carried out in terms of Accuracy (A), Detection Rate (DR) and False positive Rate (FPR) the following equations:

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

True positive (TP): Classifying an intrusion as an intrusion.
False positive (FP): Incorrectly classifying normal data as an intrusion .
True negative (TN): Correctly classifying normal data as normal.
False negative (FN): Incorrectly classifying an intrusion as normal.

Table presents the Root Mean Square Error (RMSE) and RMSE presents the difference between the actual and the desired outputs based on the confusion matrix. The model which has lower RMSE is a more efficient than a model having a larger RMSE. Meanwhile ROC value calculated based on true positive and false positive. The large value of ROC indicates the ability of a model to detect intrusions while the lower value present the weakness of a model.

ROC (Receiver Operating Characteristics)
AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model . The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis. ROC value calculated based on true positive and false positive. The large value of ROC indicates the ability of a model to detect intrusions while the lower value present the weakness of a model

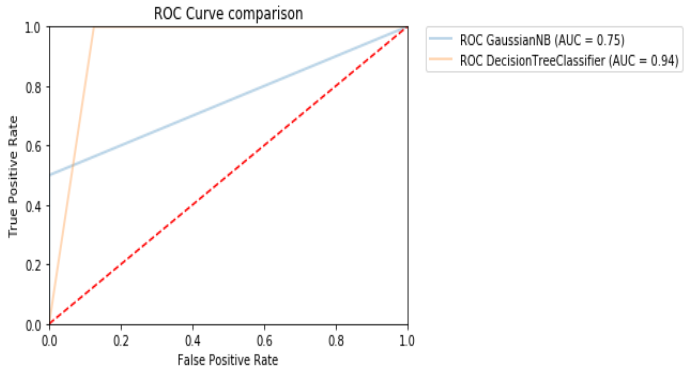

Figure :3 ROC curve comparison

5. CONCLUSION

In this paper Intrusion detection is considered as a classification problem where each record can be classified into normal or a particular kind of intrusion. Intrusion detection using machine learning have attracted more and more interests in recent years. As an important application of machine learning ,an accurate intrusion detection model is built by choosing an effective classification approach. This paper shows the comparison of the most well known classification algorithms like C4.5 decision trees and Naive Bayes has been carried out . These algorithms are tested with the KDD data-set. Effective classifier is identified by comparing the performances based on the accuracy and confusion matrix. Performance calculation is done by considering only the important attributes for the intrusion detection. From the observed results it can be concluded that the C4.5 decision trees classifier outperforms other classifiers for the considered data-set and parameters. It has the accuracy of 99%.

REFERENCES

[1]. M. Almseidin, M. Alzubi, S. Kovacs and M. Alkasassbeh, "Evaluation of machine learning algorithms for intrusion detection system," 2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, 2017, pp. 000277-000282.

[2]. Arul, Amudha & Subburathinam, Karthik & Sivakumari, S. (2013). Classification Techniques for Intrusion Detection An Overview. International Journal of Computer Applications. 76. 33-40. 10.5120/13334-0928.

[3]. Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection

[4]. Kanagalakshmi.R, V. Naveenantony Raj," Network Intrusion Detection Using Hidden Naïve Bayes Multiclass Classifier Model," International Journal of Science, Technology & Management ,Volume No.03, Issue No. 12, December 2014.

[5]. M. Alkasassbeh, G. Al-Naymat et.al," Detecting Distributed Denial of Service Attacks Using Data Mining Technique," (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, pp. 436-445, 2016. Science and Information Technologies, Vol. 6 (2), pp. 1096-1099, 2015.

[6]. Jasreena Kaur Bains ,Kiran Kumar Kaki ,Kapil Sharma," Intrusion Detection System with Multi-Layer using Bayesian Networks", International Journal of Computer Applications (0975 – 8887) Volume 67– No.5, April 2013.

[7]. Dewan Md. Farid, Nouria Harbi, Mohammad Zahidur Rahman , Combining Naive Bayes and Decision Tree for Adaptive Intrusion Detection, Proc. of Intl. Journal of Network Security & Its Applications (IJNSA), Volume 2, Number 2, 2010, pp.12-25.

[8]. Domingos P. and Pazzani M., Beyond Independence: Conditions for the optimality of the simple Bayesian Classifier, In proceedings of the 13th Intnl. Conference on Machine Learning, 1996, pp.105-110.

[9]. V. Hema and C. Emilin Shyni, " DoS Attack Detection Based on Naive Bayes Classifier, " Middle-East Journal of Scientific Research 23 (Sensing, Signal Processing and Security): 398-405, 2015.

[10]. Yi-Chi Wu, Huei-Ru Tseng, Wu Yang* and Rong-Hong Jan," DDoS detection and trackback with decision tree and grey relational analysis", Int. J. Ad

Hoc and Ubiquitous Computing, Vol. 7, No. 2, 2011.

[11]. Dewan Md. Farid, Nouria Harbi, Emna Bahri, Mohammad Zahid ur Rahman, Chowdhury Mofizur Rahman," Attacks Classification in Adaptive Intrusion Detection using Decision Tree "International Journal of Computer, Electrical, Automation, Control and Information Engineering, Vol:4, No:3, 2010.

[12]. Quinlan, C4.5: Programs for Machine Learning, 1993, Morgan Kaufmann Publishers, San Mateo, CA.

[13]. Sabhnani M, Serpen G(2003), Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context, In Proc. of the Intl. Conference on Machine Learning, Models, Technologies and Applications, vol. 1, pp. 209–215.

[14]. Gharibian F, Ghorbani A.A , Comparative Study of Supervised Machine Learning Techniques for Intrusion Detection, Proc. of the Fifth Annual Conference on Communication Networks and Services Research, 2007, pp. 350–358.

[15]. Ohta S, R. Kurebayashi and K. Kobayashi. , Minimizing false positives of a decision tree classifier for intrusion detection on the internet, Journal of Networks System Management, vol.16, 2008, pp.399–419. ISSN 1064-7570.