ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING
Vol. 2, Issue 3, March 2014

Intrusion Detection Using Random Naives Bayes Classifier In Smart Grids

Purnima.N[1], Omprakash P[2]

PG Student (M.E Communication Systems), Electronics and Communication Engineering,

Velammal College of Engineering and Technology, Madurai, TamilNadu, India[1]

Assistant Professor, Electronics and Communication Engineering,

Velammal College of Engineering and Technology, Madurai, TamilNadu, India[2]

Abstract: Smart grids (SG) represent succeeding step in modernizing this electrical grid. The communications network is combined with the Smart grid so as to collect data that may be used to increase the potency of the grid, reduce power consumption, and improve the reliability of services, among different varied benefits. Smart Grid communication networks are distinctive in their giant scale. . The Wireless networks in communication setting are going to be exposed to several threats, in order that SGDIDS can realize attacks using Random Forest Naives Bayes Classifier. Random Forest Naives Bayes is trained using information that's relevant to their level and additionally improves detection. This paper proposes a FPGA primarily based network intrusion detection in communication network of smart Grid to detect and classify malicious data and possible cyber attacks.

Keywords: Smart Grids; Wimax; Random Naives Bayes Classifier; Cyber security attacks; FPGA; Advanced Metering Infrastructure

I. INTRODUCTION

When the legacy power infrastructure is augmented by a communication infrastructure, it becomes a smart grid. This additional communication infrastructure facilitates the exchange of state and control information among different components of the power infrastructure. As a result, the power grid can operate more reliably and efficiently. Although deploying the smart grid enjoys enormous social, environmental and technical benefits, the incorporation of information and communication technologies into the power infrastructure will introduce many security challenges.

For example, it is estimated that the data to be collected by the smart grid will be an order of magnitude more than that of existing electrical power systems. This increase in data collection can possibly introduce security and privacy risks. Moreover, the smart grid will be collecting new types of information that were not recorded in the past, and this can lead to more privacy issues. All the essential parts of the smart grid will be its communication networks. The transmission system is located at the power plant and the control centres of Neighbourhood Area Network (NAN). Each NAN comprises a number of Building Area Networks (BANs) and provides them interfaces to the utility's wide-area network. Here, BANs are customer networks and belong to the second tier of the shown system. Each BAN consists of a number of networks, Home Area Network (HANs). The HAN is a customer premises network which manages the on-demand power requirements of end users. There is no standard definition of the networks. The different components of the power infrastructure of the smart grid are networked together to exchange information, there is a potential increase of the security risk of the system. For example, it will increase the complexity of the electrical power grid, which in turn can increase new security vulnerabilities. Also, the number of entry points that can be used to gain access to the electrical power system will increase when all of the components are networked together.

II. SMART GRID

Smart Grid is a set of technologies that integrate modern information technologies with present power grid system. Along with many other benefits, two-way communication, updating users about their consuming behaviour, controlling home appliances and other smart components remotely, and monitoring power grid's stability are unique features of Smart Grid. To facilitate such kinds of novel features, Smart Grid needs to incorporate many new devices and services. For communicating monitoring, and controlling of these devices/services, there may also be a need for many new protocols and standards. However, the combination of all these new devices, services, protocols, and standards make Smart Grid a very complex system that is vulnerable to increased security threats—like any other complex systems are. In particular, because of its bidirectional, interoperable, and software-oriented nature, Smart Grid is very prone to cyber attacks.

If proper security measures are not taken, a cyber attack on Smart Grid can potentially bring about a huge catastrophic impact on the whole grid and, thus, to the society. Thus, cyber security in Smart Grid is treated as one of the vital issues by the National Institute of Standards and Technology and the Federal Energy Regulatory Commission.

Fig 1. Representation of Wimax based communication in Smart Grid

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING
Vol. 2, Issue 3, March 2014

III. ADVANCED METERING INFRASTRUCTURE

AMI serves as a bridge for providing bidirectional information flow between user domain and utility domain. AMI's main functionalities encompass power measurement facilities, assisting adaptive power pricing and demand side management, providing self-healing ability, and interfaces for other systems. AMI is usually composed of three major types of components, namely, smart meter, data concentrator, and central system and bidirectional communication networks among those components. Being a complex system in itself, AMI is exposed to various security threats such as privacy breach, energy theft, illegal monetary gain, and other malicious activities. As AMI is directly related to revenue earning, customer power consumption, privacy and secures smart grid's infrastructure.

IV. COMMUNICATION TECHNOLOGIES IN SMART GRID

There are several technologies under consideration for the communication network of SGs. They can be classified into wired or wireless technologies. Among the technologies proposed for Smart Grids is ZigBee, due to its capabilities for real-time monitoring of multiple targets as well as self-organization, self-configuration, and self-healing. This makes it for example suitable to home area networks (HANs) applications. ZigBee uses the unlicensed 2.4-GHz band for wireless communications, which makes it vulnerable to interference from other technologies. In addition, ZigBee uses low data rates and the devices typically have limited memory. Alternatively, Wi-Fi has sufficient bandwidth foremost applications but is a power-hungry standard. On the other hand, there are several technologies under consideration for the neighbourhood area network (NAN) and NAN-to-NAN (N2N) parts of the Smart Grid, where data is sent from houses to utility centres. Among the wireless standards considered for SGs are cellular networks (i.e., UMTS or LTE) and wireless mesh networks (WMNs) such as Wimax. Both types of technologies have large bandwidth and are capable of supporting various QoS requirements, making them suitable for most Smart Grid applications. WMNs are capable of self-organization, self-configuration, and self-healing and can transmit using multihopping. An alternative to wireless networks is to use the wired Power line Communications (PLC). There are several advantages for using the power grid as a communication medium. First, power cables readily exist, even in rural areas, and thus the cost of implementing the communication network is low. Second, power cables are owned by the utility company, which provide a degree of independence from other networks. On the other hand, there are some disadvantages for using PLC. Devices that are turned off cannot be reached, which is an aspect that has to be considered by routing protocols. Moreover, communication signals that are transmitted through PLC channels experience attenuation. There are mainly two types of PLC communications: narrowband PLC (NBPLC), and broadband PLC (BPLC). Only data rates in the order of a few tens of kbps can be achieved with NBPLC, while BPLC can achieve data rates up to 100 Mbps. However, BPLC suffers from frequency-selective fading over large distances. PLC can be used in HAN parts of the Smart Grid where it can provide a reliable alternative in case of failure in wireless links.

V. SECURITY IN SMART GRID

In response to growing concerns about the current electric grid's vulnerability to cyber attack, there's a push on several fronts to develop better security solutions for the Smart Grid. Today, utilities typically link grid monitoring and control systems to open networks such as the Internet and critics charge they are not doing enough to reduce risks. Watch this section for insights on the future of cyber security – as well as solutions for safeguarding physical properties such as substations and transmission towers. Recognizing and acting on the vulnerabilities of the smart grids is crucial to avoid rushing headlong into a national security threat that energy utilities do not understand and cannot prevent. The grid's greatest problem is to protect software and hardware against intentional, accidental and natural threats ranging from hackers to electromagnetic pulses (EMPs). There is a perception that cyber security threats against the grid will not happen right away. However, there were a number of sophisticated attacks. A successful attack against the grid would affect far more people than an attack against a company. Smart grids are thought of as isolated networks protected by firewalls and requiring a VPN for remote access, but that kind of perimeter security is not sufficient any more. Smart meters that aim to revolutionize electric power may also be the grid's biggest security flaw. That's because the meters often rely on inherently insecure wireless networks.

VI. INTRUSION DETECTION USING RANDOM NAYES BAYES CLASSIFIER

Random forests are an ensemble learning method for classification (and regression) that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. The method combines Breiman's "bagging" idea and the random selection of features in order to construct a collection of decision trees with controlled variation. The selection of a random subset of features is an example of the random subspace method is a way to implement stochastic discrimination. It is better to think of random forests as a framework rather than as a particular model. The framework consists of several interchangeable parts which can be mixed and matched to create a large number of particular models, all built around the same central theme. The simplest type of decision to make at each node is to apply a threshold to a single dimension of the input. This is a very common choice and leads to trees that partition the space into hyper-rectangular regions. However, other decision shapes, such as uses linear or quadratic decisions are also possible. Predictors determine how a prediction is made for a point, given that it falls in a particular cell of the space partition defined by the tree. Simple and common choices here include using a histogram for real valued outputs, or constant predictors for categorical data.

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING
Vol. 2, Issue 3, March 2014

In principle there is no restriction on the type of predictor that can be used, for example one could fit a Support Vector Machine or a spline in each leaf; however, in practice this is uncommon. If the tree is large then each leaf may have very few points making it difficult to fit complex models; also, the tree growing procedure itself may be complicated if it is difficult to compute the splitting objective based on a complex leaf model. However, many of the more exotic generalizations of random forests, e.g. to density or manifold estimation, rely on replacing the constant leaf model. The splitting objective is a function which is used to rank candidate splits of a leaf as the tree is being grown. This is commonly based on an impurity measure, such as the information gain or the Gini gain. The method for injecting randomness into each tree is the component of the random forests framework which affords the most freedom to model designers. Breiman's original algorithm achieves this in two ways:

1. Each tree is trained on a bootstrapped sample of the original data set.
2. Each time a leaf is split, only a randomly chosen subset of the dimensions are considered for splitting.

In Breiman's model, once the dimensions are chosen the splitting objective is evaluated at every possible split point in each dimension and the best is chosen. This can be contrasted with the method of Criminisi, which performs no bootstrapping or sub sampling of the data between trees, but uses a different approach for choosing the decisions in each node. Their model selects entire decisions at random (e.g. a dimension threshold pair rather than a dimension). The optimization in the node is performed over a fixed number of these randomly selected decisions, rather than over every possible decision involving some fixed set of dimensions.

A. Breiman's Algorithm

Each tree is constructed using the following algorithm:

1. Let the number of training cases be *N*, and the number of variables in the classifier be *M*.
2. We are told the number *m* of input variables to be used to determine the decision at a node of the tree; *m* should be much less than *M*.
3. Choose a training set for this tree by choosing *n* times with replacement from all *N* available training cases (i.e., take a bootrap sample). Use the rest of the cases to estimate the error of the tree, by predicting their classes.
4. For each node of the tree, randomly choose *m* (out of *M*) variables on which to search for the best split. Calculate the best split based on these *m* variables in the training set. Base the decision at that node using the best split.
5. Each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier).

For prediction a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in. This procedure is iterated over all trees in the ensemble, and the mode vote of all trees is reported as the random forest prediction.

B. Relationship to Nearest Neighbors

Given a set of training data

$$D_n = \{(X_i, Y_i)\}_{i=1}^{n}$$

a weighted neighbourhood scheme makes a prediction for a query point X, by computing

$$\hat{Y} = \sum_{i=1}^{n} W_i(X) Y_i$$

for some set of non-negative weights $\{W_i(X)\}_{i=1}^{n}$. The set of points X_i where $W_i(X) > 0$ are called the neighbours of X. A common example of a weighted neighbourhood scheme is the K-NN algorithm which sets $W_i(X) = \frac{1}{K}$ if X_i is among the K closest points to X in D_n and 0 otherwise.

Random forests with constant leaf predictors can be interpreted as a weighted neighbourhood scheme in the following way. Given a forest of M trees, the prediction that the m-th tree makes for X can be written as

$$T_m(X) = \sum_{i=1}^{n} W_{im}(X) Y_i$$

Where $W_{im}(X)$ is equal to $1/K_m$ if X and X_i are in the same leaf in the m-th tree and 0 otherwise, and K_m is the number of training data which fall in the same leaf as X in the m-th tree. The prediction of the whole forest is

$$F(X) = \frac{1}{M} \sum_{m=1}^{M} T_m(X) = \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{n} W_{im}(X) Y_i$$

$$F(X) = \sum_{i=1}^{n} \left(\frac{1}{M} \sum_{m=1}^{M} W_{im}(X)\right)$$

Which shows that the random forest prediction is a weighted average of the Y_i 's, with weights

$$W_i(X) = \frac{1}{M} \sum_{m=1}^{M} W_{im}(X)$$

The neighbours of X in this interpretation are the points X_i which fall in the same leaf as X in at least one tree of the forest. In this way, the neighbourhood of X depends in a complex way on the structure of the trees, and thus on the structure of the training set.

C. Packet Normalization

Pkt_norm (Packet Normalization) can be shredded in transit, and processing these shards of data can increase system load. The system at one end of the connection might have a habit of braking up transmitted packets into tiny bits or perhaps some router in the middle thinks that our packets are too large and splits them up to digest them more easily. So, instead of nice whole packets our network will receive small bits or fragments. of the trained dataset, then they will display as normal packets.

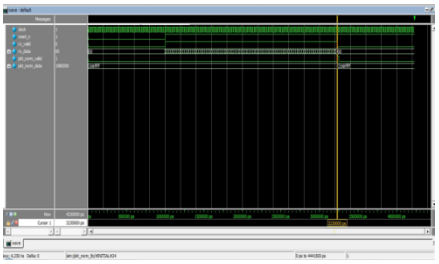

Fig 2. Representation of packet normalization

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING
Vol. 2, Issue 3, March 2014

Packet normalization is done here for the TCP and IPV4 protocols. The UDP and IPV4 can also be done depends upon the packets. In the Packet Normalization the TCP and IPv4 data fed in the backend program. So that the packets arriving later will be compared with the Random Naives Bayes train dataset. If the incoming packet does not match the packets of the trained dataset, then they will display as attacked packets. If the incoming packet matches the packets, they will display as normal packets. Output waveform for the packet normalization coding is obtained in Questasim software of VHDL and which is represented in below figure 2.

D. Training Data

Random Naïve Bayes training set we are fixing the values according to the substation of the Smart Grids. The values of the data differ for each substation.

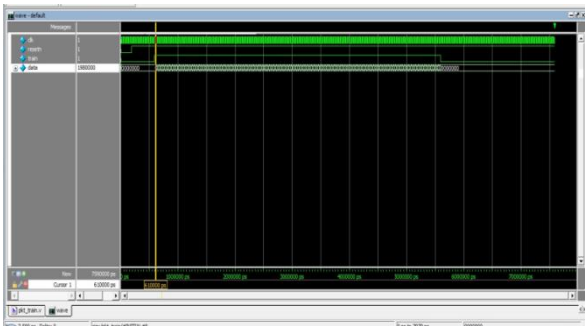

Fig 3. Representation of training data using Random Naives Bayes Classifier

After finding the values for each substation we are given those values as the training data set to the training algorithm of the Random Naïve Bayes Classifier. The incoming packets may belong to any type of class i.e it may be a attacked packet or may be a normal incoming packets. Form the training dataset the Random Naives Bayes classifier will test the incoming packets. Output waveform obtained in the questasim_6.4c for the training data is represented in the below figure 3 to find the attacked packets in Smart Grids.

E. Attacked Packets

If the incoming packet does not match the packets of the trained dataset of the Random Naïve Bayes Algorithm, then output will display as attacked packets. Output waveform for the attacked packets is obtained in Questasim software of VHDL and which is represented in below figure 4.

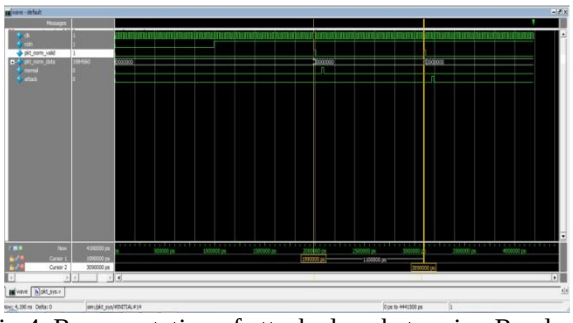

Fig 4. Representation of attacked packets using Random Naives Bayes Classifier

VII. CONCLUSION

The Wireless mesh networks in communication environments of Smart Grids are secured from many threats, since SGDIDS found the attacks using Random Naives Bayes Classifier. The attacked packets are found using Breiman's Algorithm of the data according to the Random Naïve Bayes Classifier which differ for each substation in the Smart Grids. Hardware behaviour will directly reflect because of using VHDL based Smart Grids Network intrusion Detection (SGDIDS). The FPGA based intrusion detection produces faster and accurate method for finding the attacked packets in the Wimax communication network of Smart Grid.

REFERNCES

[1] S. X. Wu and W. Banzhaf, "The use of computational intelligence in intrusion detection systems: A review," Appl. Soft Comput., vol. 10, no. 1, pp. 1–35, 2010.
[2] [Mentor] Mentor Graphics Company
[3] Ruiz-Llata, G. Guarnizo, and M. Y?benes- Calvino. FPGA implementation of a support vector machine for classification and regression. WCCI 2010 IEEE World Congress on Computational Intelligence, IJCNN, july 2010
[4] [IEC61850] International Electrotechnical Comission (IEC) 61850.J. Cho, B. Benson, S. Cheamanukul, and R. Kastner. Increased performance of FPGA-based color classification system. Annual IEEE Symposium on Field-Programmable Custom Computing Machines, pages 29–32, 2010.M.
[5] D. Kim and J. Park, "Network based intrusion detection with support vector machines," in Proc. ICOIN, 2003, vol. 2662, Lecture Notes in Computer Science, pp. 747–756.
[6] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn.,vol. 20, pp. 273–297, 1995 [Online].Available: http://www.springerlink.com/index/K238JX04HM87J80G.pdf
[7] S. R. Gunn, "Support vector machines for classification and regression,"Faculty Eng., Sci., Math. School Electron. Comput. Sci.,Tech.Rep.,May1998[Online].Available:http://pubs.rsc.org/en/Content/ArticlePDF/2010/AN/B918972F/ 2009-12-23

BIOGRAPHIES

Purnima N, received bachelor of engineering in Electronics and Communication from Kalasalingam University, Virudhunagar, India in 2012. Currently doing Master of Engineering in Communication Systems at Velammal College of Engineering and Technology, Madurai, TamilNadu, India. Research interests include Network and Security. At present, she is engaged in Smart Grid applications.

Omprakash P, received Bachelor of Engineering in Electronics and Communication Engineering from PSR Engineering College, Sivakasi, TamilNadu, India in 2004. Worked as a Team leader in Reliance Communications, Madurai, TamilNadu, India. Received Master of Engineering in Communication Systems from Sri Krishna College of Engineering and Technology, Coimbatore, TamilNadu, India in 2011. Currently working as a Assistant Professor in Velammal College of Engineering and Technology, Madurai, TamilNadu, India.