An Intrusion Detection System Based on Big Data for Power System

SichengZeng[1 2, a]

[1] Globe Energy Interconnection Research Institute, Beijing 102209, China

[2] China Electric Power Research Institute, Beijing 100192, China

[a]orgazeng@163.com

Keywords: Power System, CPS, Data Mining, Intrusion Detection

Abstract. On the background of information and energy interconnection, the whole power system generated a huge amount of data with diverse structure, complicated sources and large scale from both cyber devices and physical components, which is a typical cyber-physical system (CPS). These data exhibit data feature such as large quantity, complicated data item, complex processing logic, long storage cycle and high frequency calculation. Therefore, from a CPS perspective, the power system is facing intrusions that are more damaging, complicated and wide spreading. Currently, most power system network intrusion detection systems are founded manually. Especially, the detection knowledge used for identify intrusion action is provided by security expert and complied into the network intrusion detection system(IDS). The defect of this approach is that it needs the continuing input of upgraded knowledge concerning the intrusion detection, which may not suit for the complex power CPS. Therefore, the expansion and adaptability of such term is not suitable in the context of big data problem. In this paper, we propose hierarchic IDS that combines misuse detection and abnormal detection for Power System. Data mining algorithms are used to build the rules by studying and analyzing historical monitor date. The prototype implemented proves that the model proposed can detect cyber-attacks accurately with low false positive and false negative rate.

Introduction

On the background of information and energy interconnection, the whole electric system generated a huge amount of data with diverse structure, complicated sources and large scale from both cyber devices and physical components, which is a typical cyber-physical system (CPS). These data exhibit data feature such as large quantity, complicated data item, complex processing logic, long storage cycle and high frequency calculation. Therefore, from a CPS perspective, the electric system is facing intrusions that are more damaging, complicated and wide spreading. The attack on Ukrainian electric system on December 23rd, 2015 demonstrates the impact of intruding electric and communication system and attacking power grid by hackers. Currently, most electric system network intrusion detection systems are founded manually. Especially, the detection knowledge used for identify intrusion action is provided by security expert and complied into the network intrusion detection system. The defect of this approach is that it needs the continuing input of upgraded knowledge concerning the intrusion detection.Therefore, the expansion and adaptability of such term is so confined that it cannot meet the needs of handlingbigamounts of data acquainted from power CPS. This article begins with the brief introduction the principles of intrusion detection system and its basic type, and then explains in detail the algorithm and realization of the big data calculation applied in the model of this article. Finally, it proposes in-depth design of the model and conduct simulation. The result of the simulation demonstrates that the intrusion detection model based on big data can effectively enhance the security of the whole power CPS.

Electric System Intrusion Detection System Introduction

Intrusion Detection System Introduction (IDS) is a dynamic security technique, which can detect and report unauthorized or abnormal behavior in the network in time. It is an important supplement to traditional security system and can exert crucial effect concerning network and system protection.

Data Source Based Classification. Host Intrusion Detection System (HIDS) is a system that conducts detection and analysis based on system log and assessment record from the hosts of intellectual convert stations and various remote supervisions and dispatch station. It usually has Proxy Detection on protected host and detects attack from continuous monitoring and analyzing system log and assessment record.Network Intrusion Detection System (NIDS) uses the original network data package of the electric system as the data source. It is usually stationed in critical network segment. The system consistently monitoring various data package in the network segment and analyze the feature of each data package or the suspicious ones. If the data package is a (or potential) intrusion package, the system will send alarm, even abort the system network directly.Hybrid Data Source Intrusion Detection System (HDSIDS) sets various of data source as detection as target in order to enhance the effectiveness of the IDS. The IDS of HDSIDS can be configured into distributed mode. Normally it installs monitor module on serve and network path that need to be supervised, reports and uploads evidence to managing server respectively and provides trans-platform intrusion detection solutions. Compared to the previous two IDS, HDSIDS has more comprehensive detection capacity. It integrates feature of the previous two IDS, therefore can either detect the attack behavior in the network flow or discover abnormality among system log.

Technique Based Classification. From a technical perspective, attacks can be classified into two categories: one is featured attack, which performs conventional attacks on known weakness of the system, the other one is abnormality attack. Accordingly, intrusion detection has also two types: misuse detection (a.k.a. feature detection) and abnormality detection (a.k.a. behavior detection).Misuse detection expresses the behavior of the attack as a mode and it aims at checking whether the behavior of the subject conform to the mode. Since the attacking mode is also known as feature, misuse detection is also called feature detection. Among current commercial product, the most common misuse detection form is to define each attack mode as an independent feature and build an attack feature date base. The difficult lies in how to design a system that can express attack phenomenon meanwhile exclude normal activities. The advantage of such method is its low misreport rate and detection of know attack. However, the efficiency of such method depends on the completeness of the detection data base.[1] Therefore, the data based has to be updated on time. Furthermore, this method cannot discover unknown attack.Different from misuse detection, abnormality detection sets the subject as the target and assumes that the attacking behavior differs from normal activity of the subject. According to this concept, this method builds the normal mode of the subject and compares it with current activities of the subject. When the examined activity disobeys the normal mode, such activity will be identified as attack. Abnormality Detection collects historical data of the normal operation from a certain period, builds the pattern of the normal behavior of the host or network connection, a.k.a. normal mode. The disadvantage of such method is how to build such mode and design algorithm in order to exclude normal operation and prevent overlooking the attack.The advantage of such method is that it does not rely on the feature of the attack but the subject of the supervision. But problems include how to build abnormality index and how to define the normal range to lower the misreport rate still have to be solved.Hybrid Intrusion Detection System integrates the previous two mode thereby produce more reliable strategy. As it is already known, misuse detection can effectively identify unknown attack and has low misreport rate. But misuse detection can hardly recognize new attack type, therefore it can ignore attacks. Meanwhile abnormality detection has a high misreport rate, although it can identify unknown attack. Hybrid Intrusion Detection System combines the advantage of both method, conducting analysis on misuse detection result and abnormality detection result of the subject before the taking the decision, thereby make more accurate and comprehensive judgment.

The application of data science in IDS

The foundation of supporting strategy with data is data science. As a brand new domain, data science is an interdisciplinary subject that integrates sociology, statistics, communication and computer science. Its development is inseparable from computer, internet and recent big data. Data science studies science and topic from data, collects data knowledge, research and apply data based on the

needs and feature of science and topics and provide data service. Data science has been widely applied in intrusion detection techniques. At this stage, the detection methods applied by IDS based on data science theory include nerve network, support vector machine, immune, cluster analysis, data mining, cloud computing and big data.

Establish electrical intrusion detection system model based on big data

Considering previous introduction on intrusion detection technique and data science, it is clear that intrusion detection mode based on big data environment has flexibility and can generate intrusion detection rules through intellectually analyze data. This chapter will conduct building electrical system intrusion detection model and explain the rules and functions of data processing in intrusion detection model. This model will center on data and consider the generating of intrusion detection rule as a data analysis process, in order to reduce the artificial part of IDS. And meanwhile, it applies multi-layer combination of misuse detection and abnormality detection to minimize omission and false report.

Intrusion Detection Model Architecture. Data analysis module generates rules on the basis of historic data. These rules are stored in rule library for intrusion detection judgment. The system captures packets on the network. After pre-processing, the captured packets are sent to monitoring module. The monitoring module compares the packets with misuse detection rule library and abnormally detection rule library to detect possible attacks. These detection results will be provided to firewall system of the network by being submitted to administrator or intrusion detection and firewall joint control system. [2]

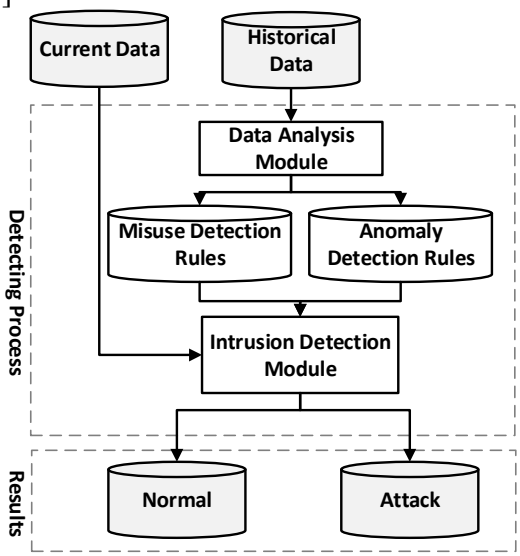

Fig. 1 Intrusion detection model architecture

Working process of data analysis module is shown as Fig. 2. It monitors and collects activity historic data of the targeted system within a period of time. After data pre-processing, historic connection record comprising a series of specific attributes is formed. Rule mining module analyzes the connection record and generates the matching rules for abnormally detection rule library. Signature generation module generates misuse detection rule library by learning signatures of training data group.

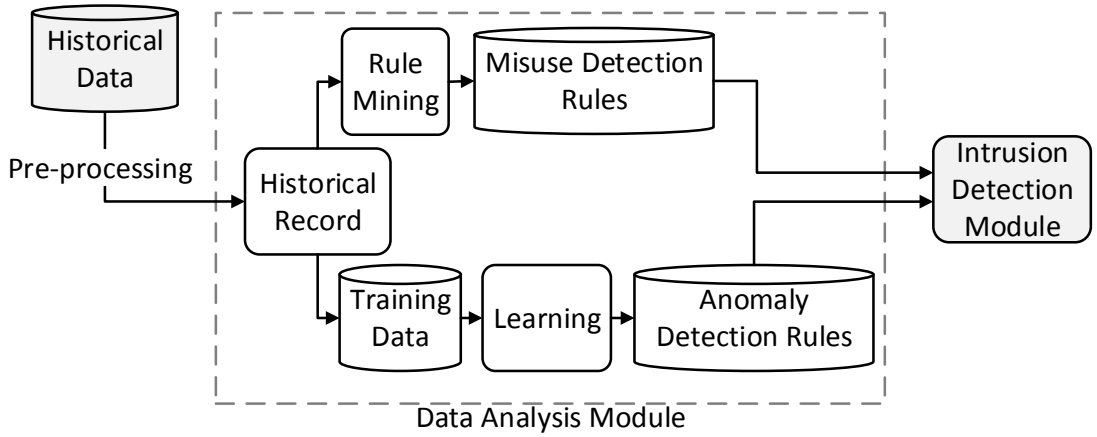

Fig. 2 Data analysis module

As described before, misuse based intrusion detection system can detect known malicious threats while abnormally based intrusion detection system will detect what deviates "normal" activity.

These two detection modules have their own advantages and disadvantages. Misuse based detection works efficiently with known malicious threat with high detection probability and low rate of false detection. However, it cannot detect unknown attack or new variety of known attack. Since the attack methods keep updating, it is almost impossible to pre-define all attack methods. Given the fact that new threats and attacks update more and more frequently, misuse based detection suffers more. Abnormally based detection can detect unknown attacks, but with high rate of false alarm. It is because that the so called "normal" or "abnormal" is relevant. Usually, they are distinguished by an abnormally threshold. If setting a low threshold for high detection sensitivity, a lot of normal behaviors that deviate a little from strict normal behaviors will be recognized as abnormal. What's more, if the traffic behaviors of attacks (such as some U2R or R2LT attacks) are very close to normal traffic behaviors, abnormally based detection can hardly detect these attacks with low rate of false detection.

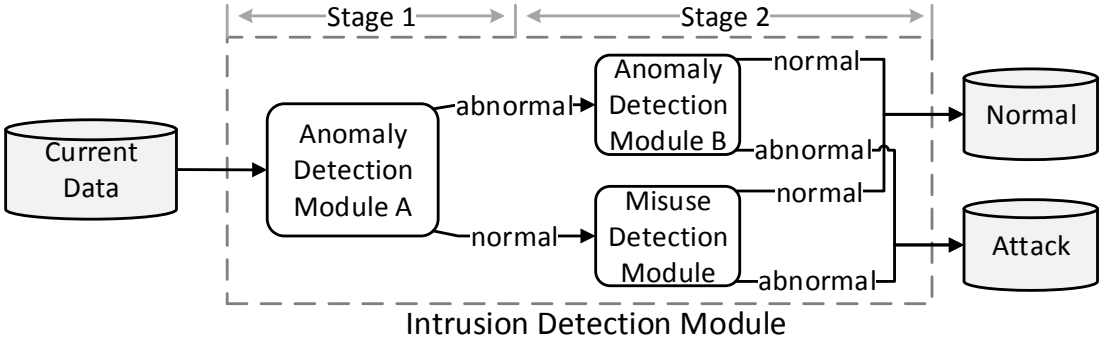

Fig. 3 Intrusion detection module

Because both misuse based and abnormally based detection modules have their own strong points, we hope to combine these two kinds of detection mechanisms and design a new mix based intrusion detection system.

The new intrusion detection module is shown below.

Algorithm Design

Algorithm of Abnormally Based Detection. The state set of classical Hidden Markov Model (HMM) is constant [3]. This influences the ability of HMM modeling random signals, and also limits the performance of classifier based on HMM.

When deploying to abnormally based detection module, we make proper adjustment to HMM. Since Bayes classifier performs well for KDD Cup99 sequence set [4], we combine advantages of HMM and simple Bayes classifier to construct multi-dimension HMM abnormally based detection module and algorithm. Thus, the number of states can match automatically to the number of real

hidden states. This enables a more accurate modeling of random signals and provides more information of structure.

To process sequence set with temporal relationship, we utilize method of matching support vector machine and sliding window. For a known data set $X= \{X_1,X_2,\ldots,X_T\}$, the size of window is w. Obtain data of length was the new data block each time. After getting a new data block, obtain the next data block by shifting 1 unit right. The length of test sequence is T, then the short sequence set contains T－w+1 sliding window short sequences with w blocks of data. The sequence set is shown as below.

$$X=\{\{X_1,X_2,\ldots,X_w\},\{X_2,X_3,\ldots,X_{w+1}\},\ldots,\{X_{T-w+1},X_{T-w+2},\ldots,X_T\}\} \quad (1)$$

In the observation sequence of length N, the average value of probability of appearing observation sequence of length w is

$$A(N)=\frac{\sum_{i=1}^{N-w-1}\left(\prod_{j=i}^{i+w-1}\frac{b_0(X_j)}{\max\limits_{1\le k\le M} b_0(k)}\right)}{N-w+1} \quad (2)$$

Therefore, we can derive the following equation [5][6]

$$A(N)=\frac{\sum_{i=1}^{N-w-1}\left(\prod_{j=i}^{i+w-1}\frac{b_0(X_j)}{\max b_0(k)}\right)}{N-w+1}=\frac{N-w}{N-w+1}A(N-1)+\frac{\prod_{j=i}^{i+w-1}\frac{b_0(X_j)}{\max\limits_{1\le k\le M} b_0(k)}}{N-w+1} \quad (3)$$

Initial value $A(w)=\prod_{i=1}^{w} b_0(X_i)$, calculate when $N\ge w$. For the situation $N<w$, since w is small, we think that there would be no abnormal behaviors under this situation. Compare the output probability of each short sequence in the short sequence set with the preset output probability threshold, all short sequences that smaller than threshold are defined to be mismatch. Then, count the number of mismatch sequences. Define $\delta=c_e/c$ as the abnormal parameter. c_e is the number of mismatch short sequences, and c is the total number of short sequences. Finally, we compare δ with threshold δ_a. If $\delta>\delta_a$, the test sequence is abnormal, and vice versa.

Different threshold values will be set for the different functions of abnormally based detection module in the model to achieve the optimum performance.

Algorithm of Misuse Based Detection. Apriori algorithm is an algorithm for frequent item set mining and association rule learning over transactional databases. It was proposed by R. Agrawal and R. Srikant in 1994. This algorithm utilize the future knowledge of frequent item set. The main idea is to use breadth-first search[7]. It generates candidate item sets of length k+1 from item sets of length k.

Firstly, scan the database to count every item, and filter out the minimum support threshold min_sup that satisfy frequent item set. Thus, we can get frequent 1-item set, L_1. Then, associate L_1 with itself to generate candidate frequent item set. Again, according to min_sup to filter away non-frequent items to generate frequent 2-item set, L_2. Use the same methodology to L_2 to generate frequent 3-item set, L_3. Keep calculating until cannot find frequent k-item set. [8][9]

ALGORITHM: APRIORI ALGORITHM

INPUT: TRANSACTIONAL DATABASE D, MIN_SUP
OUTPUT: FREUQENT ITEM SET L IN D

```
1:  Scan D to get frequent 1-item set, L1.
2:  for(k=2;Lk-1≠∅;k++){
3:      Ck=aproiri_gen(Lk-1,min_sup);
4:      for t in D begin
5:          Ct=subset(Ck,t);
6:              for c in Ct
7:              c·count++;
8:          end
9:      Lk={c∈Ck|c·count≥min_sup}
10: }
```

PROCEDURE APRIORI_GEN(L_{K-1},MIN_SUP)


```
//CANDIDATE SET GENERATION FUNCTION
1:   for every l1 ∈ Lk-1 begin
2:       for every l2 ∈ Lk-1 begin
3:         if(l1[1]==l2[1]∧l1[1]==l2[2]∧…∧l1[k-2]==l2[k-2])∧l1[k-1]<,
4:             c={l1[1],l1[1],…,l1[1],l1[1],l1[1]};
5:                For every s⊂c begin
6:                  if(s∉Lk-1) then
7:                          delete c;
8:                      else add c to Ck;
9:                end
10:           end
11:      end
12:  end
```

Experiment and Analysis. Experiment data utilize KDD Cup99 as training and detection data. To simplify processing procedure, only use part data set of whole KDD Cup99 (about 10%, including training data set and test data set). Ever data is comprised of 41 characteristic attributes and 1 decision attribute. To ensure efficiency of execution, randomly select 100 thousand records as training data set and test data set in the experiment and divide them into 10 groups randomly.

Every TCP/IP connection has 41 attributes and is labeled its type (e.g. normal or specific attack method). This data set contains multiple simulation attacks in the network environment. It mainly can be divided into five types as shown below.

Table 1Types of data set

Attack type	Type description	Example
Normal	Normal request	
Dos	Denial of service attack	TCP flood attack
Prob	Probe system vulnerability	Port scan
R2L	Unauthorized access to remote server	Password guess
U2R	Unauthorized use of Root authority	Buffer overflow

During the experiment, validate detection rate, false alarm rate, and failure alarm rate respectively. The equations for all above three is shown below.

$$\text{Detection rate}=\frac{\text{Number of detection}}{\text{Number of abnormal behaviors}}\times 100\% \quad (4)$$

$$\text{False alarm rate}=\frac{\text{Number of false alarms}}{\text{Number of normal behaviors}}\times 100\% \quad (5)$$

$$\text{Failure alarm rate}=\frac{\text{Number of failure alarms}}{\text{Number of total attacks}}\times 100\% \quad (6)$$

Result in Table 2 indicates that the detection performance of the whole system is stable and efficient to detect inrusion behaviors in the network data.

Table 2 Data validation result

Detection rate	False alarm rate	Failure alarm rate
95.71%	0.44%	0.21%

Conclusion

Misuse detection methods was based on the judgment of behavior by signature database, the detection accuracy is high, but this method cannot detect unknown attacks; anomaly detection method can detect known attacks and unknown variants of attacks, but the rate of false positives and false negative rate was higher than former method. In this article, a method that combines a multi-level anomaly detection and misuse detection was proposed. The method was specifically adapted to the characteristics of large data of Power CPS, thus it can not only effectively detect known attacks, but also has the ability to detect unknown attacks. At the same time the system can reduce the dependence

of artificial and experience in intrusion detection system by detection rules generated from historical data analysis and mining. Experimental results showed that the modified model can effectively enhance the detection rate and reduce false positives, false negative rate.

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (61471328).

References

[1] RUAN Yaoping，Yi Jiangbo,ZHAO Zhansheng.The intrusion detection model and method of computer system [J].Computer Engineering，1999,25(9):63-65.

[2] WANG Lihui,LI Tao,ZHANG Xiaoping. A network intrusion detection system firewall [J]. Application Research Of Computers，2006，（03）：95-97.

[3] XIE Yi, YU Shun-zheng. A dynamic anomaly detection model for web user behavior based on HSMM[C]. International Conference of CSCWD '06, Guangzhou, CHINA: IEEE Press, 2006: 451- 460.

[4] AMOR N B, BENFERHAT S, ELOUEDI Z. Naive Bayes vs decision trees in intrusion detection systems[c]. Proceedings of the ACM Symposium on Applied Computing, Nicosia, Cyprus:ACM，SIGAPP，2004: 420- 424.

[5] FTANZ P. Bayesian network classifiers versus selective formula not Shown-NN classifier [J]. Pattern Recognition， 2005, 38(1) : 1- 10.

[6] CHEBMLU S, ABRAHAM A, THOMAS J P. Feature deduction and ensemble design of intrusion detection systems [J]. Computers and Security, Elsevier, Amsterdam. 2005, 24(4) : 295- 307.

[7] RAJASEGARAR S, LECKIE C, PALANISWAMI M. Centered hyper ellipsoidal support vector machine basedanomaly detection[C]. IEEE International Conferenceon Communications. United States: Institute of Electricaland Electronics Engineers Inc,2008: 1610–1614.

[8] HATHAWAY R J, BEZDEK J C, HUBAND J M. Scalable visual assessment of cluster tendency for large datasets [J]. Pattern Recogn, 2006, 39:1315–1324.

[9] FALLAH S, TRICHLER D, BEYENE J. Estimating number of clusters based on a general similarity matrix with application to microarray data [J]. Statist. Appl. Genet. Mol. Biol,2008，7(1):1- 23.

[10] SHI Hong-bo, WANG Zhi-hai, HUANG Hou-kuan, et al. A restricted double level Bayesian classification model [J].Journal of Software, 2004, 15(2):193-199

