XVII International Conference on Recent trends in Engineering, Science and Management (ICRTESM-19)
Mahratta Chamber of Commerce, Industries and Agriculture, Tilak Road, Pune (India)
28th July 2019 www.conferenceworld.in ISBN : 978-93-87793-99-6

K-Nearest Neighbor Method for Privacy Preserving in Data Mining

Niranjana Garg[1] Dr. Amit Sharma[2]

[1]M.Tech Scholar Department of Computer Science & Engineering
Vedant college of Engineering and Technology Bundi Affiliated From R.T.U.Kota, India
[2]Associate Professor Department of Computer Science & Engineering
Vedant college of Engineering and Technology Bundi Affiliated From R.T.U. Kota, India

Abstract

An intrusion is defined as any activity being performed in a system which might result in triggering such an event that compromises the security of the system. Intrusion Detection Systems (IDSs) use the statistical analysis methods for detecting any anomalies such that an activity can be differentiated as either normal or malicious. Machine learning is known as an Artificial Intelligence based technology using which the programs can be learned and the data patterns can be identified. Machine learning is used to explore the algorithms that can perform learning and perform data prediction. The two categories among which the machine learning algorithms are broadly categorized are supervised learning and unsupervised learning. To perform intrusion detection, SVM classification method was applied previously. To perform intrusion detection KNN classifier is applied by the proposed research work. On KDD dataset, the proposed and existing methods are implemented. In terms of accuracy, the results of both the techniques are tested. The outcomes show that the intrusion detection system provides the best outputs when KNN classifier is used.

KEYWORDS: IDS, KDD, KNN, SVM

I Introduction

Over the years, the demand of intrusion detection system has increased as with each day the information being stored and processed is increasing. By monitoring the surroundings of applications huge amount of data is generated by the networking systems [1]. The devices detect any kinds of suspecting behaviors from the surroundings. An intruder can cause any kind of vulnerability in the computer network due to which the users can be attacked. An intrusion is known as an activity that results in tampering such an event due to which the security of system can be compromised. A kind of alarm is caused by an

ISBN : 978-93-87793-99-6

intrusion detection system and it is possible to identify the violations of system. The systems can be alerted in case of any false messages, videos or mails. An intrusion detection system is a tool that is used as a guard for securing the systems against any kinds of intrusions [2]. Any malicious activities which cannot be identified by a common firewall can be detected through IDS. In any computer system, the regions, computer applications and sensitive services can be attacked by unauthorized users. The computer applications can face the data driven attacks. The intrusions can face network attacks in sensitive services and also the sensitive files can be accessed by unauthorized logins [3]. In the recently occurring incidents and activities, providing traditional model is not feasible for detecting intrusions. To identify any kind of attack, the network is analyzed manually or few fixed abnormal patterns are provided when applying traditional models. Recently, it is easy to access the policy due to which the network traffic can be improved with the help if internet and threats of attacking are identified. These activities can help in improving the network analyst and it also becomes difficult to detect the intrusions. Highly dynamic efficient methods are needed for automating the intrusion detection process. From these systems, learning can be changes and any kinds of intrusions existing in these systems can be detected. Pattern matching is known as the process through which intrusions are detected by performing comparisons with the known attack signatures [4]. This technique helps in generating the signatures from audit records and comparing against the current activities such that intrusions can be detected. The attack signatures in which common binary patterns are included can be included to identify abnormal activity. Data mining techniques help in extracting the interesting factors hidden in the database. To collectively form the knowledge few relationships, classes and patterns can be identified. The data mining methods can be applied to process the huge amount of data instantly. Few of the techniques are based on human intervention such that the intrusions can be detected. Many days or weeks of time can be consumed to detect the new signatures of intrusions. Spending days or weeks in infeasible to identify an intrusion since with each day, the network traffic is increasing. Machine learning is a kind of artificial intelligence that performs learning in programs and identifies the data patterns [5]. Machine learning explores the algorithms using which learning can be performed and data can be predicted. They are commonly known as machine learning algorithms. Learning is important for machine learning algorithms before making any data predictions. Learning helps an algorithm to show the examples of data and correct predictions. It is important to include the amounts of examples in the range of several thousands [6]. Once machine learning algorithm performs learning, it is possible to perform predictions on data. For example, machine learning helps in monitoring the heart related patients in hospitals. Machine learning algorithm can be applied in the learning phase to show the heart rate of a patient and the current time. To determine if the heart rate of patient is normal or not, the predicted heart rate and real heart rate are compared. The two commonly used machine learning algorithms are supervised and unsupervised learning [7]. A classifier using which the simplest classifications can be performed is known as KNN which is also commonly known as a non-parametric supervised learning algorithm. There is no assumption included within the underlying data distribution. Based on the closest training samples of feature space, the samples are classified. Naive Bayes Algorithm is an algorithm using which a classification method is applied on the basis of

Bayes Theorem in which the independence among predictors is assumed. Based on the classifier's assumption, the presence of particular feature in a class in unrelated to the availability of another feature.

II Literature Review

Altyeb Altaher, (2017) presented a hybrid approach to classify websites as Legitimate, Suspicious, or Phishing. In order to develop this hybrid technique, the proposed algorithm used two stages. The hybrid approach used two classification models called KNN and SVM [8]. In the first phase, KNN approach was implemented. This algorithm was quite efficient and strong to the noisy data. One more robust classification model called SVM was implemented in the second phase. After integrating the simplicity of KNN approach, the proposed approach enhanced the efficiency of SVM classifier. Various simulation tests were performed to evaluate the proposed algorithm. The obtained results depicted that the proposed algorithm showed the maximum accuracy rate of 90.04% as compared to other existing algorithms.

Jayshree Jha, et.al (2013) presented a novel research work on the basis of two significant parts. The first part reviewed the attack detection with the help of SVM algorithm along with the other approaches presented by different researchers [9]. Moreover, in the second part, a new technique was presented for selecting optimum feature for detecting attack. A hybrid algorithm was presented to choose the related features. This algorithm fused the filter and wrapper models. The size of the database was decreased to improve the performance and discovery accurateness of a discovery model based on SVM algorithm. In addition, the training and testing time could be reduced by reducing the feature set.

L.Dhanabal, et.al (2016) used KSL-KDD dataset for performance analysis. The efficiency of various classification algorithms was studied to identify the irregularities present in the patterns of network traffic [10]. The association of protocols obtainable in frequently utilized network protocol stack was examined to produce irregular network traffic. The network protocol stacks was studied with the intrusions launched by attackers. These attacks produced the abnormal network traffic. The classification algorithm was used along with WEKA software for performance analysis. This work revealed numerous facts tied amid the protocols and network intrusions.

Wathiq Laftah Al-Yaseen, et.al (2015) introduced a multi-level hybrid model called IDS. In this work, the support vector machine and extreme learning machine were used to successfully identify known and unknown attacks [11]. An advanced k-means clustering algorithm was also proposed in this work to improve the performance of classification models. This clustering algorithm constructed an optimum training dataset. This algorithm integrated the novel small training datasets. These datasets defined entire real training dataset. Thus, the proposed approach decreased the training time of classification models. The proposed model performed better as compared to other techniques designed and applied on similar dataset to detect intrusion. In addition, the propose approach also showed good performance in terms of accuracy than all analyzed algorithms.

Amol Borkar, et.al (2017) reviewed Internal-IDS and IDS models. The data mining and forensic algorithms based on real time were implemented in these models [12]. A variety of data mining methodologies were proposed for cyber investigation to help in attack recognition. This work presented different techniques to detect attack on the basis of several analyses

provided by different researchers. The review provided in this work proved helpful to draw the conclusion. The use of proposed approach enhanced the accurateness and discovery rate up to 95%. On the other hand, the existing techniques provided about 90% of accurateness and discovery rate. Thus, these results clearly indicated that the proposed approach performed better as compared to other existing algorithms in terms of accuracy and intrusion detection.

Jianguo Yu, et.al (2018) analyzed that the recognition of attack in the rail transit field was the main aim of data security model. An expert attack recognition system called BAS was designed to detect attack and mis-operation of subway environment control subsystem. Moreover, the knowledge base and inference engine design were developed in the expert system as well [13]. The expert systems were utilized to detect mis-operation and mis-use attack. In addition, the black and white list rules were added to avoid irregular attack. The rules provided support to protect the data security of subway environment control system to large extent. This technique also provided data security to multiple subsystems of subway. At present, this system is just being utilized in investigative state because of some flaws. Nevertheless, the IDSs can be implemented to the entire metro area by using big data theory.

III Research Methodology

The network traffic classification approach is applied for categorizing the data traffic as malicious or non-malicious. The malicious activities of active clients are predicted by this method. To classify the network by applying proposed approach, three important steps are applied. To cluster the data as similar or dissimilar, k-means clustering approach is applied. To refine the specified dataset as input, few issues such as redundancy and missing values are removed. To calculate the central point of network, the k-means clustering technique is implemented. The arithmetic mean of overall dataset is calculated in this step. From the central point Euclidian distance is calculated for differentiating similar and dissimilar points. Similar data points are included in one cluster and others in separate clusters. To categorize the data points into two different classes, the SVM classification model is implemented in the last stage. To improve the accuracy and performance of classification method, the non-clustered data points are also clustered by applying KNN classification model. The Euclidian distance is calculated and similar and dissimilar kind of data is differentiated by calculating Euclidian distance.

ISBN : 978-93-87793-99-6

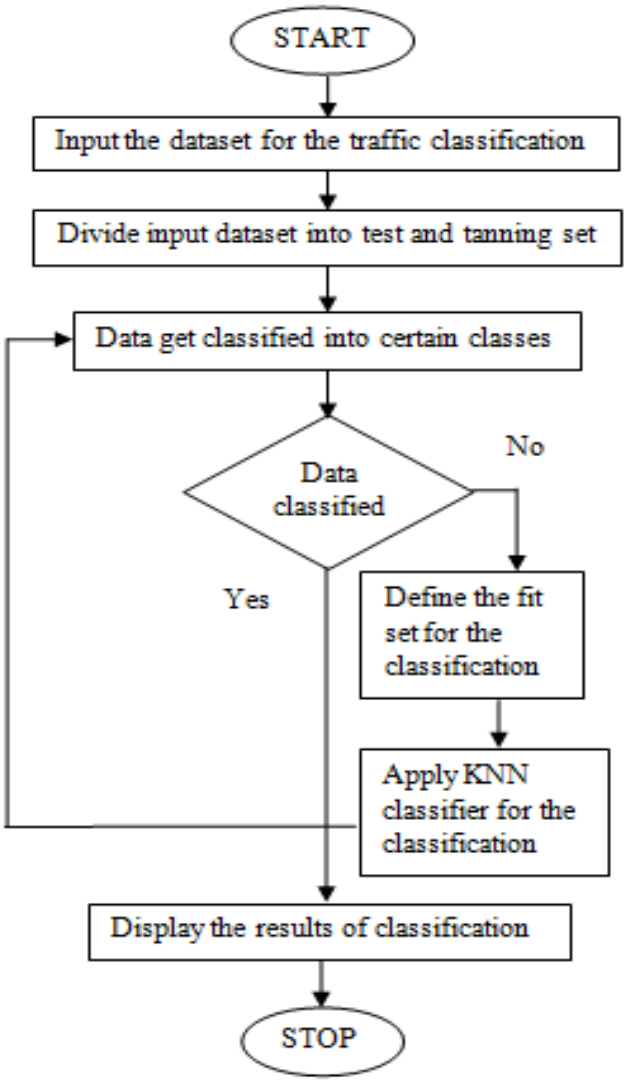

Figure 1: Proposed Flowchart

IV Experimental Results

The proposed research is implemented in Python and the results are evaluated by comparing proposed and existing methods in terms of different performance parameters.

ISBN : 978-93-87793-99-6

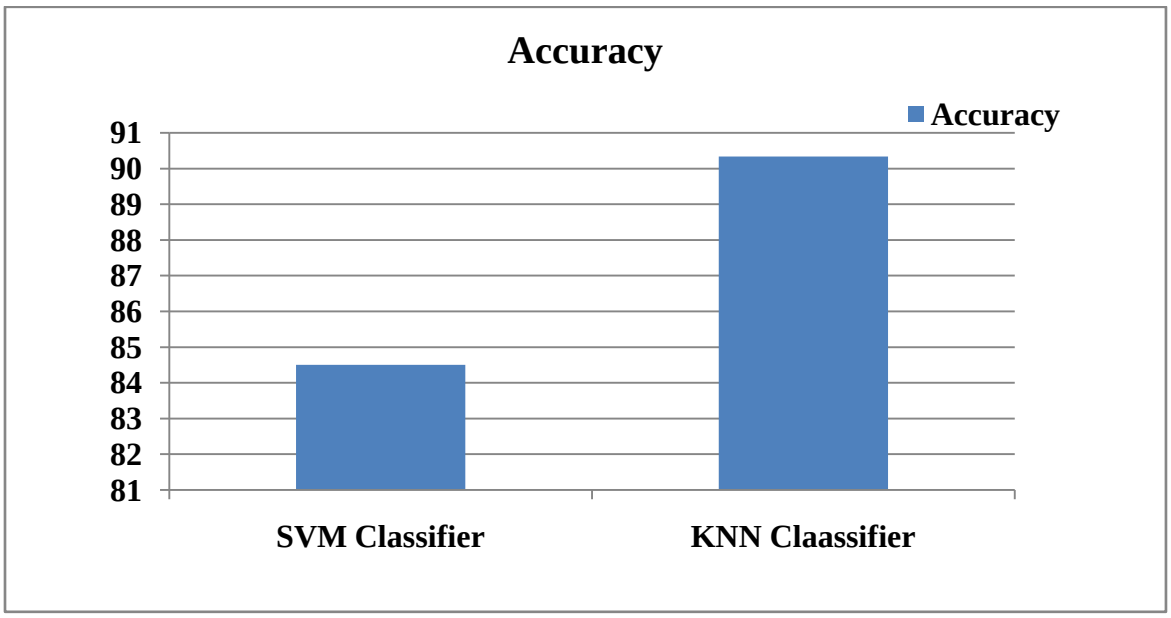

Fig 2: Accuracy Comparison

A comparative analysis of performances of SVM and KNN is shows in figure 2. The outcomes of comparison graph show that accuracy level of KNN classifier is better than SVM classifier.

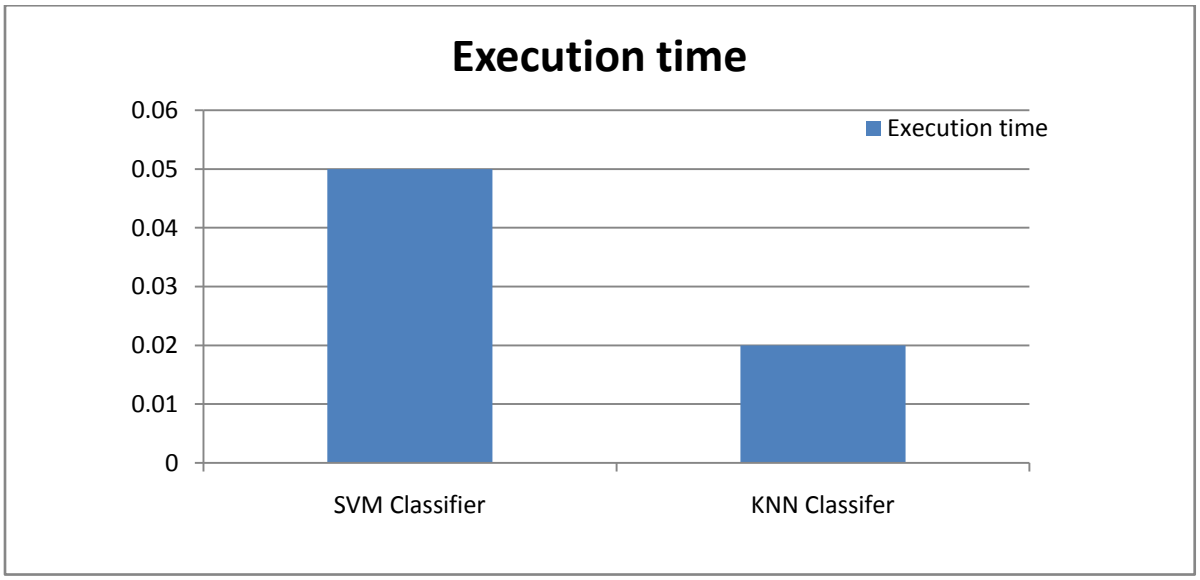

Fig 3: Execution Time

Based on execution time, the performances of proposed and existing algorithms are compared as show in figure 3. As shown in the comparison graph, in terms of execution time, the results of KNN approach are better as compared to SVM.

ISBN : 978-93-87793-99-6

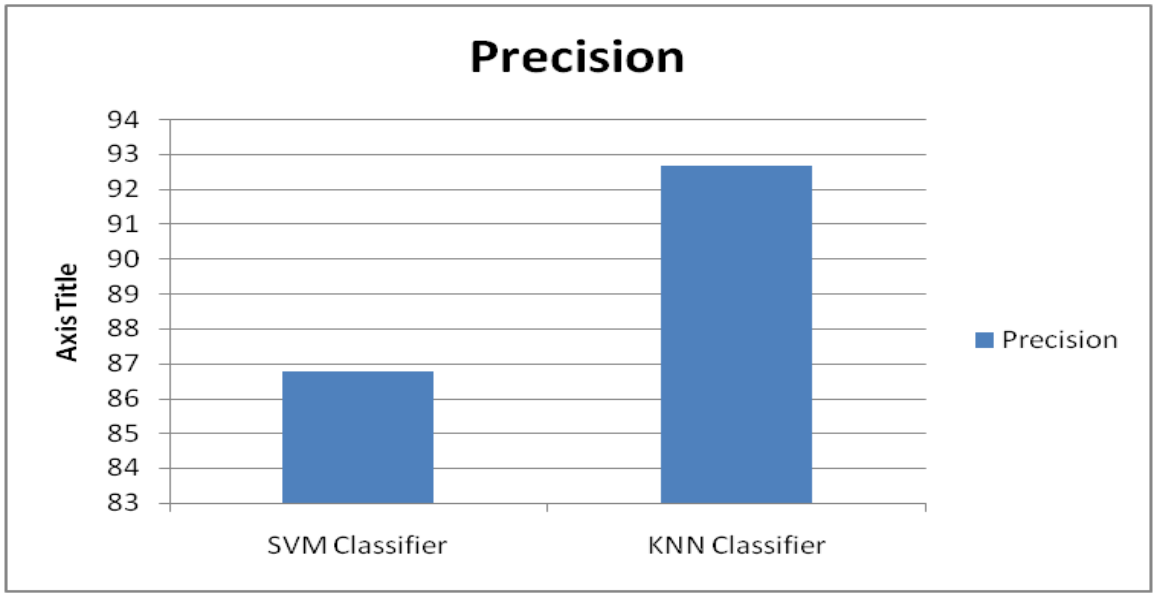

Fig 4: Precision Analysis

A comparative analysis of performances of SVM and KNN is shows in figure 4. The outcomes of comparison graph show that precision level of KNN classifier is better than SVM classifier.

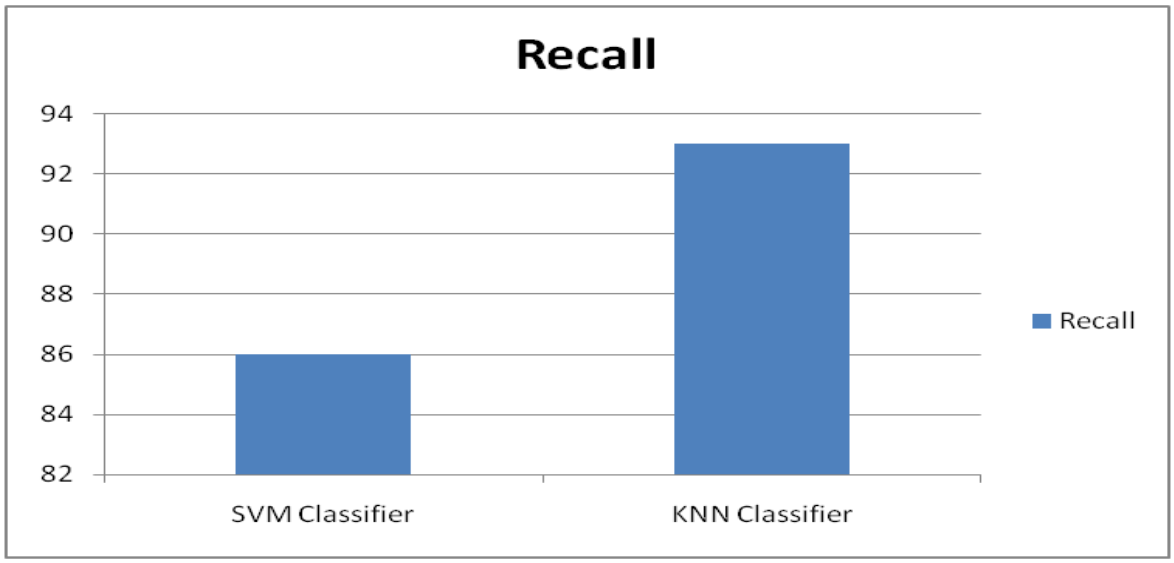

Fig 5: Recall Analysis

A comparative analysis of performances of SVM and KNN is shows in figure 5. The outcomes of comparison graph show that recall level of KNN classifier is better than SVM classifier.

V Conclusion

The host-based IDSs are the systems that monitor the devices on which they are installed. For executing the monitoring program the states of main system are monitored from the audit logs to the program execution. This research aims to study the different intrusion detection techniques that are adaptive, highly dynamic and that can be applied in huge network traffic. Based on the closest training samples of feature space, the samples are classified. Along with the labels of training images, the feature vectors are stored such that they can be used in training process. To perform labeling of k-nearest neighbors, the unlabelled question point is ruled out. Python simulator is used to implement the proposed methodology. With respect to accuracy and execution time, the result evaluations are performed. The outcomes show that in comparison to SVM classifier, the KNN classifier provides better outputs. The accuracy is improved from 5 to 8% by applying KNN classifier.

VI References

[1] Mamadou Alpha Barry, James K. Tamgno, Claude Lishou, ModouBambaCissé, "QoS Impact on Multimedia Traffic Load (IPTV, RoIP, KDD) in Best Effort Mode", International Conference on Advanced Communications Technology(ICACT), 2018

[2] Ahmed Fawzy Gad, "Comparison of Signaling and Media Approaches to Detect KDD SPIT Attack", IEEE, 2018

[3] Mario A. Ramirez-Reyna, S. Lirio Castellanos-Lopez, Mario E. Rivero-Angeles, "Connection Admission Control Strategy for Wireless KDD Networks Using Different Codecs and/or Codec Mode-sets", The 20th International Symposium on Wireless Personal Multimedia Communications (WPMC2017)

[4] MurizahKassim, Ruhani Ab. Rahman, Mohamad AzraiA.Aziz, Azlina Idris, Mat IkramYusof, "Performance Analysis of KDD over 3G and 4G LTE Network", IEEE, 2017

[5] Jan Holu, Michael Wallbaumy, Noah Smithy and HakobAvetisyan, "Analysis of the Dependency of Call Duration on the Quality of KDD Calls", IEEE, 2018

[6] Mohammad Tariq Meeran, Paul Annus, Yannick Le Moullec, "Approaches for Improving KDDQoS in WMNs", IEEE, 2017

[7] JanuszHenryk Klink, Tadeus Uh, "Quality-aware network dimensioning for the KDD service", IEEE, 2017

[8] Altyeb Altaher, "Phishing Websites Classification using Hybrid SVM and KNN Approach", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 6, 2017

[9] Jayshree Jha, Leena Ragha, "Intrusion Detection System using Support Vector Machine", 2013, International Journal of Applied Information Systems (IJAIS), Foundation of Computer Science FCS, New York, USA International Conference & workshop on Advanced Computing

[10] L.Dhanabal, Dr. S.P. Shantharajah, "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms", 2016, International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 6

[11] Wathiq Laftah Al-Yaseen, Zulaiha Ali Othmana, Mohd Zakree Ahmad Nazri, "Multi-Level Hybrid Support Vector Machine and Extreme Learning Machine Based on Modified K-means for Intrusion Detection System", 2015, Expert Systems With Applications

[12] Amol Borkar ; Akshay Donode ; Anjali Kumari, "A survey on Intrusion Detection System (IDS) and Internal Intrusion Detection and protection system (IIDPS)", 2017 International Conference on Inventive Computing and Informatics (ICICI), Pages: 949 – 953

[13] Jianguo Yu, Pei Tian, Haonan Feng, Yan Xiao, "Research and Design of Subway BAS Intrusion Detection Expert System", 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Pages: 152 – 156