NKB-S: Network Intrusion Detection Based on SMOTE Sample Generation

Yuhan Suo[1], Rui Wang[2], Senchun Chai[1(✉)], Runqi Chai[1], and Mengwei Su[1]

[1] School of Automation, Beijing Institute of Technology, Beijing 100081, People's Republic of China
{yuhan.suo,chaisc97,r.chai,3220180645}@bit.edu.cn
[2] Institute of Automation, Chinese Academy of Sciences, Beijing 100190, People's Republic of China
rwang5212@ia.ac.cn

Abstract. This paper mainly studies the problem of sample generation for imbalanced intrusion datasets. The NKB-SMOTE algorithm is proposed based on the SMOTE algorithm by combining the K-means algorithm and using a mixture of oversampling and undersampling methods. The Synthetic Minority Oversampling (SMOTE) Technique sample generation is performed on the minority class samples in the boundary cluster, the Tomek links method is used for the majority class samples in the boundary cluster to undersample the boundary cluster, and the NearMiss-2 method is used to undersample the overall data. Then, multi-classification experiments are conducted on the UNSW-NB15 dataset, and the results show that the proposed NKB-SMOTE algorithm can improve the generation quality of samples and alleviate the fuzzy class boundary problem compared with the traditional SMOTE algorithm. Finally, the actual experiment also verifies the effectiveness of the intrusion detection model based on NKB-SMOTE in real scenarios.

Keywords: Network security · Intrusion detection · Machine learning · Sample generation · NBK-SMOTE

1 Introduction

With the development of science and technology, the Internet is constantly integrating with our work, study and life, and has become an indispensable part of human society [1]. However, new cybersecurity threats are emerging, such as the 2016 cyberattack on an energy company in Ukraine that caused widespread power outages in Kyiv, and the ransomware spread rapidly around the world in 2017, causing huge losses to people's information security and property security [2,3]. Therefore, ensuring network security has also become a necessary link.

Industrial system intrusion detection is an important part of network security protection. At present, the research of intrusion detection mainly includes

This work was supported by the Basic Research Program under Grant JCKY******* B029.

F. Sun et al. (Eds.): ICCSIP 2022, CCIS 1787, pp. 130–147, 2023.
https://doi.org/10.1007/978-981-99-0617-8_10

four aspects, namely statistical analysis, data mining, feature engineering and machine learning. The intrusion detection technology based on statistical analysis is based on the analysis of network behavior by statistical means to determine whether the system is under attack, including regularization entropy based on mixed standard deviation [4], multivariate model based on attribute feature correlation [5], Marko based Intrusion detection is carried out using methods such as the husband process [6]. Intrusion detection based on data mining is based on the correlation between a large number of data to find the law between data [7], the most common method is the intrusion detection model based on K-means clustering algorithm [8], it has been proved that classification analysis methods such as mapping data feature attributes into different types can improve the detection effect. Intrusion detection based on feature engineering mainly includes feature dimensionality reduction and feature selection. Feature dimensionality reduction can extract the same features of attributes to reduce the data dimension [9], and feature selection can select the most effective features from a series of features [10]. Both methods can improve the detection efficiency to a certain extent. Intrusion detection based on machine learning is an emerging intrusion detection method in recent years [11,12], which has higher detection accuracy than traditional methods. It can be seen that most of the intrusion detection algorithms are based on data. However, because the industrial intrusion dataset is usually unbalanced, the classification effect of the minority class samples is often poor. In the actual system, the harm caused by the minority class samples is often greater, so how to improve the detection effect of the minority class samples remains to be studied.

In order to improve the classification effect of the classification model on the minority data in the imbalanced data set, researchers have carried out a lot of research. The main method is to balance the dataset by resampling before training the model. The methods of resampling mainly include undersampling, oversampling and mixed sampling. The undersampling algorithm balances the dataset by reducing the majority class samples in the data, the oversampling algorithm balances the dataset by increasing the minority class samples in the data, and the mixed sampling combines the characteristics of undersampling and oversampling. The simplest undersampling algorithm is simple random undersampling, which randomly deletes the majority class sample data, but may lose important information; Bo et al. proposed an undersampling algorithm based on majority class classification, which eliminates the randomness of deleting samples [13]; Lin et al. cluster the majority class samples, and then delete the samples [14]; Padmaja et al. perform undersampling based on the filtering method, which improves the adaptive ability of sampling [15]. The oversampling algorithm may lead to overfitting due to the invalid replication of minority class samples. In order to avoid the occurrence of overfitting, Kang et al. [16] proposed an oversampling method based on K-nearest neighbors, and Chawla et al. [17] proposed a random oversampling-based SMOTE algorithm, which is also one of the classic algorithms in oversampling algorithms. Some scholars proposed improved algorithms based on the SMOTE algorithm, which greatly improving the effect of the original SMOTE algorithm [18–21]. Mixed sampling is mainly a combination

of the above two sampling methods. Seiffert et al. proposed a mixed sampling method that combines random undersampling and random oversampling [22]. It is proved by experiments that the mixed sampling method has better classification effect than single sampling. Li et al. proposed a hybrid sampling method combining the distance-based undersampling method and the improved SMOTE algorithm to improve the accuracy of data classification [23].

The research on the processing of unbalanced intrusion datasets helps to improve the detection of intrusion detection, especially for a small number of samples, and is of great practical importance for industrial system security problems. Therefore, this paper investigates the sample generation problem of unbalanced intrusion dataset, and the innovation points are shown as follows.

- This paper improves the SMOTE algorithm to make up for the problems of fuzzy class boundary and uneven distribution of generated samples in the original algorithm.
- The proposed NKB-SMOTE algorithm not only improves the quality of the newly generated samples, but also enables the data to be balanced faster.
- The classification experiments for the existing dataset and the actual industrial control system dataset illustrate the effectiveness of the proposed algorithm compared to the traditional algorithm.

In the second section, the relevant knowledge used in this paper is introduced. In the third section, the NKB-SMOTE sample generation algorithm based on the SMOTE technology is proposed to generate samples. Finally, machine learning multi-classification experiments are carried out on the UNSW-NB15 dataset and the dataset obtained from real scenes, which further verifies the effectiveness of the algorithm proposed in this paper.

2 Related Knowledge

2.1 Intrusion Detection System

Intrusion Detection (ID) [24]: By monitoring the working network system, it can detect intrusion attempts, stop intrusion behaviors, and find intrusion results to ensure network system security. By recording network flow information and analyzing information characteristics, the attack behavior and abnormal operation can be judged. Essentially, ID is a classification operation on characteristic data such as network flow and system operation logs.

Intrusion Detection System (IDS) [25]: It is a defense system that protects network data, classifying and identifying network flow data in real time, and proactively issuing alarms and taking defensive measures. The basic structure of IDS is shown in the Fig. 1, and the main contents are as follows.

- Event generator: obtain network data stream and provide current data stream events to other parts of IDS;
- Event analyzer: analyze the network data flow and pass the analysis results to the responder;

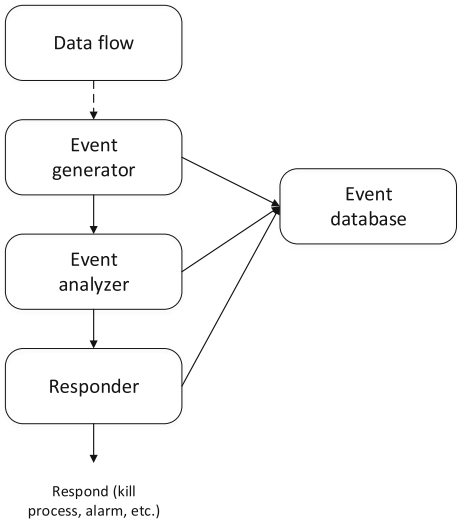


Fig. 1. The schematic diagram of the basic structure of the intrusion detection system.

- Responder: react to the event analysis result;
- Event database: store intermediate data and final data.

Intrusion detection systems can be divided into the following three categories according to the data source:

- Host-based IDS: Its data sources are mainly log files (application logs, router logs, system logs, etc.), port usage records, and host audit information. Detection is performed through user usage records, but the detection is often delayed, that is, after the host is attacked, the intrusion is detected and alarm information is given.
- Network-based IDS: Its data source is mainly network packet data packets in the monitored network. IDS obtains data packets from the network card (wired or wireless) connected to the network device, and detects the attack behavior in the network through feature matching analysis.
- IDS based on mixed data sources: its data sources include the data sources of the first two IDSs, including audit records from hosts and data packets in the monitored network. The IDS based on mixed data sources is usually distributed, and can simultaneously discover the attack behavior inside the host and the abnormal behavior in the detected network.

2.2 SMOTE Technology

The full name of SMOTE technology is Synthetic Minority Oversampling Technique. Unlike random oversampling algorithm, SMOTE technology is no longer limited to simple copying of minority class samples, but by inserting new samples between minority class sample points and their adjacent minority class sample

points, which avoids the occurrence of model overfitting problems that may be caused by random oversampling techniques, as shown in Fig. 2.

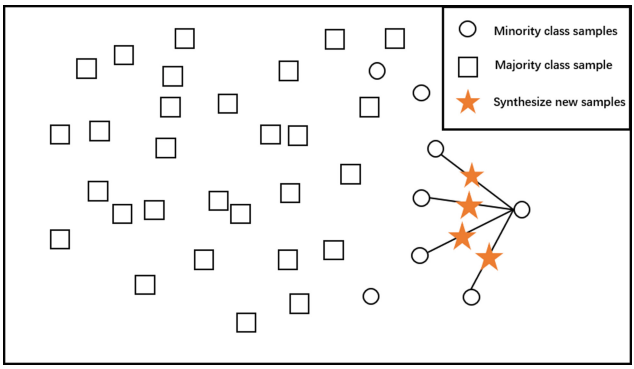


Fig. 2. The schematic diagram of new sample points generated by SMOTE technology.

Assuming that the number of minority class samples in an imbalanced dataset is T, the expected number of minority class samples to increase is NT. Among them, N is the sample multiple to increase, (usually an integer). For each minority class sample point x_i, $i \in 1, 2, ..., T$, K-nearest minority class sample points ($k \geq N$) are found by Euclidean distance.

Then, N sample points $(x_{near1}, x_{near2}, ..., x_{nearN})$ are randomly selected from them, and the minority class sample points x_i and each x_{near} are brought into the interpolation formula (1). In this way, the interpolation sample point x_{new} is obtained, which is the newly generated minority class sample point. By performing the above operations on T minority class sample points, NT new minority class sample points can be obtained to achieve the purpose of balancing the dataset.

$$x_{new} = x_i + RAND\,(0, 1) * (x_{near} - x_i)\,, \tag{1}$$

where $RAND(0, 1)$ is a number randomly generated in the interval $[0, 1]$.

2.3 Evaluation Metrics

It is well known that one of the important evaluation metrics of intrusion detection performance is accuracy, which indicates the ratio of the number of correctly classified samples to the total number of samples, and one drawback of accuracy rate is that it does not perfectly reflect the effect of the model. Therefore, the metric of precision rate is proposed. The precision indicates the proportion of the number of correctly classified positive samples to the number of all predicted positive samples, and this metric is inversely proportional to the magnitude of the false positive rate. Therefore, this paper will subsequently adopt the precision as one of the evaluation metrics.

The AUC constant is usually used to evaluate the merit of a binary classifier, and its geometric meaning is the area under the curve in the ROC curve. The Fig. 3 shows the ROC curve, whose vertical axis is the true positive rate, i.e., the ratio of the number of positive sample predicted outcomes to the actual number of positive samples, and the horizontal axis is the false positive rate, i.e., the ratio of the number of negative sample outcomes predicted to be positive to the actual number of negative samples. Obviously, the closer the value of AUC is to 1 and the closer the ROC curve is to the upper left corner of the coordinate, the better the classifier is, which is an intuitive evaluation indicator.

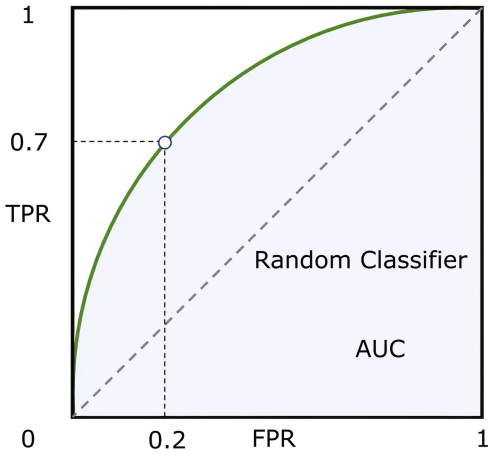


Fig. 3. The ROC curve diagram.

2.4 The Statement of Problem

The SMOTE technology analyzes the minority class samples and artificially synthesizes new samples to add to the dataset to alleviate the overfitting problem caused by simply copying the sample points, but it also has some defects:

- The quality of synthetic samples: if one of the root sample and the auxiliary sample is a noise sample, the new sample will most likely fall in the majority class area;
- Fuzzy class boundary problem: Synthesize new samples from the minority class samples at the class boundary, and the minority class samples synthesized by interpolation will also fall in the overlapping area of the two classes, thus blurring the boundaries of the two classes more;
- Minority class distribution problem: The original minority class dense area is still relatively dense after SMOTE, while the sparse distribution area is still relatively sparse.

In order to avoid the above-mentioned problems of SMOTE, this paper will study a new sample generation algorithm based on SMOTE technology, which can improve the quality of sample generation, blur the class boundary problem, and make the data reach balance faster.

3 Intrusion Detection Model Based on SMOTE Sample Generation

In order to avoid the above-mentioned problems of SMOTE technology, this section proposes the NKB-SMOTE algorithm to optimize the SMOTE technology. It not only combines the K-means algorithm and the SMOTE algorithm, but also uses a mixture of oversampling and undersampling.

3.1 The Theoretical Basis of NKB-SMOTE Algorithm

In the original SMOTE technology, during the sample generation process of the interpolation operation, it is necessary to set the number of adjacent sample points k of the minority class samples, which has a certain blindness, and the optimal value can only be obtained through repeated testing experiments. Such operations are not only time-consuming, but also cannot guarantee optimal values. The cluster center of the K-means algorithm after the clustering of the minority class samples is used as the core point of the interpolation formula, and the samples of the SMOTE technology are generated in combination with other minority class sample points in the cluster. At the same time, the majority class samples within the cluster are undersampled by Tomek links. This processing not only solves the blindness of choosing the k value of the number of adjacent samples, but also alleviates the problem of fuzzy class boundaries.

In the NKB-SMOTE algorithm, when the K-means clustering algorithm is used to cluster the minority class samples in the dataset, it should be noted that the cluster center of the cluster must be limited to the minority class sample points. After the clustering is completed, k clusters are obtained, and each cluster is classified by the idea of classifying the minority class samples using the SMOTE algorithm. As shown in Fig. 4, according to the number of minority class samples and majority class samples in the cluster, the cluster is divided into a safe cluster S, a noise cluster N, and a boundary cluster $B = (B_1, B_2, ..., B_m)$, where m represents the number of boundary clusters. For the boundary cluster, with the cluster center as the core point, combined with other minority class sample points in the cluster, new sample points are generated by interpolation formula (1), and then the majority class samples in the cluster are undersampled by Tomek links algorithm. As shown in Fig. 5, it can be seen intuitively that the newly generated sample points are always in the circle with the radius of the connecting line between the cluster center and each minority class sample point, that is, always in the minority class area, so as to avoid the problem of marginalization of newly generated sample points. In addition, the Tomek links algorithm is used to undersample the majority class sample points in the cluster, so that

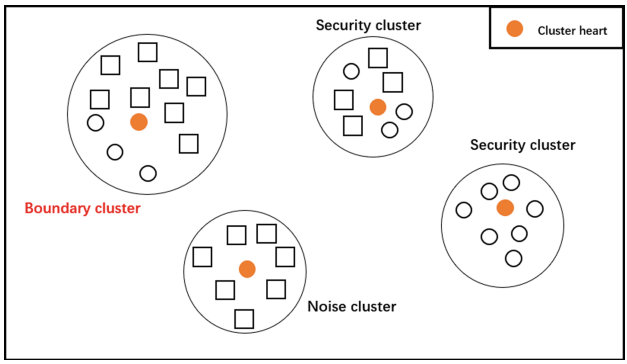


Fig. 4. The schematic diagram of clustering with minority class samples as cluster centers.

the boundary between the majority class and the minority class is clearer, which contributes to improve the classification accuracy. Finally, while the minority class samples are generated in the boundary cluster and the majority class samples are undersampled by the Tomek links algorithm, the majority class samples are undersampled by the NearMiss-2 algorithm on the entire dataset. The imbalance classification performance of the NearMiss-2 algorithm is better than that of the Tomek links algorithm, which can further speed up the data balance.

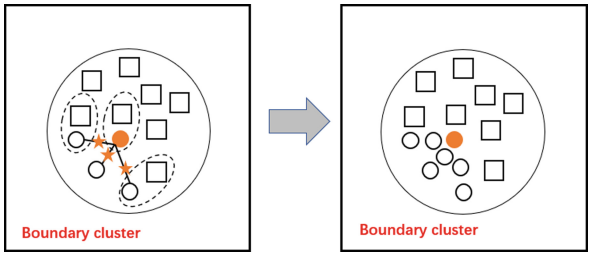


Fig. 5. The schematic illustration of oversampling and undersampling within boundary clusters.

It can be seen from the above description and schematic diagram that the NKB-SMOTE algorithm avoids the selection of the k value of the number of adjacent samples through the K-means clustering method, makes a pre-judgment for each cluster, and uses a mixture of oversampling and undersampling in the boundary clusters. The generated sample points are all in the cluster, which effectively alleviates the problem of fuzzy class boundaries in the traditional SMOTE algorithm, and the Tomek links algorithm undersampling can make the class boundaries clearer. And each minority class sample point in the boundary cluster is used only once, which is different from the repeated use of minority

class samples in the traditional SMOTE algorithm, which can effectively avoid the generation of duplicate data and meaningless data. Finally, by combining the global NearMiss-2 algorithm with undersampling of the majority class samples of the dataset, the dataset can reach balance faster.

3.2 The Research Content of NKB-SMOTE Algorithm

The focus of the research content of NKB-SMOTE algorithm is cluster classification and mixed sampling within boundary clusters. Suppose the given dataset is $X = \{x_1, x_2, ..., x_n\}$, where n represents the number of samples in the dataset. The set of minority class samples in the dataset is $X_{\min} = \{x_1, x_2, ..., x_p\}$, where p represents the number of minority class samples in the dataset. The set of samples of the majority class in the dataset is $X_{\max} = \{x_{p+1}, x_{p+2}, ..., x_n\}$, where $n - p + 1$ represents the number of majority class samples in the dataset.

K-means is performed with the sample points in the set of minority class samples X_{min} of the dataset X as the cluster center Clustering operation, after the clustering is completed, the set of clusters $C = (C_1, C_2, ..., C_k)$ is obtained, where k is the number of clusters. The sample points in each cluster may contain several minority class samples and several majority class samples. According to the quantitative relationship between the majority class samples and the minority class samples in the cluster, the type of the cluster is judged.

1. Security Cluster Judgment:
   Calculate the number of minority class samples and the number of majority class samples in clusters C_i, $i = 1, 2, ..., k$, after clustering is completed. As shown in formula (2), if the number of minority class samples is greater than the number of majority class samples (including all minority class samples and the number of majority class samples is 0), the cluster can be judged as a Security cluster.
   $$|C_i \cap X_{\min}| \geq |C_i \cap X_{\max}| . \tag{2}$$
2. Noise Cluster Judgment:
   Calculate the number of minority class samples and the number of majority class samples in clusters C_i, $i = 1, 2, ..., k$, after clustering is completed. As shown in formula (3), if the number of minority class samples is 1 (except for the cluster center, which are all majority classes), it can be known that the cluster center is a noise point, and the cluster can be judged as a noise cluster.
   $$|C_i \cap X_{\min}| = 1. \tag{3}$$
3. Boundary cluster judgment:
   Calculate the number of minority class samples and the number of majority class samples in clusters C_i, $i = 1, 2, ..., k$, after clustering is completed. As shown in formula (4), if the number of samples of the minority class is greater than 1 and less than the number of samples of the majority class, it can be known that the cluster center is the boundary point, and the cluster can be

judged as the boundary cluster $B_j \in (B_1, B_2, ..., B_m)$, where $0 < j \leq m$, where m is the number of boundary clusters.

$$1 < |C_i \cap X_{\min}| < |C_i \cap X_{\max}| . \tag{4}$$

After completing the cluster classification, it is necessary to oversample and undersample the minority class samples and majority class samples in the boundary cluster $B = (B_1, B_2, ..., B_m)$, respectively.

In the original SMOTE algorithm, each minority class sample point is used as a core point, combined with its K-nearest neighbor samples, and the samples are generated by interpolation formula. In the NKB-SMOTE algorithm, after the boundary cluster $B = (B_1, B_2, ..., B_m)$ is obtained, the cluster center of the boundary cluster is the core point, and the sample is generated by combining other minority sample points in the cluster. The interpolation formula of NKB-SMOTE algorithm is shown in formula (5).

$$x_{new} = c_i + RAND\,(0, 1) * (x_j - c_i) . \tag{5}$$

where x_{new} is a newly generated sample, c_i, $i = 1, 2, ..., m$, is the cluster center of the i-th boundary cluster, $RAND(0, 1)$ is a randomly generated number between 0 and 1, x_j, $j = 1, 2, ..., t$, represents the original minority sample points in the cluster except for the cluster center, and t is the number of original minority samples in the cluster.

After completing the classification of the clusters, the Tomek links algorithm is also required to undersample the majority class samples in the boundary cluster $B = (B_1, B_2, ..., B_m)$:

1. Calculate the distance $d(x_{min}, x_{max})$ between each minority class sample in the boundary cluster and all majority class samples in the cluster.
2. Judging the distance between each minority class sample point x_{min}, the majority class sample point x_{max} pair, whether there is a sample point y in the cluster (not limiting the majority class or minority class) satisfies any of the following formulas (6–7):

$$d\,(x_{\min}, y) < d\,(x_{\min}, x_{\max}) , \tag{6}$$

$$d\,(y, x_{\max}) < d\,(x_{\min}, d_{\max}) . \tag{7}$$

3. If there is no such y sample, the pair x_{min}, x_{max} is the Tomek Links sample pair, and the majority class samples in the Tomek Links sample pair are deleted.

3.3 Detailed Description of NKB-SMOTE Algorithm

The specific operation steps of the NKB-SMOTE algorithm are shown as follows:

Suppose the given dataset is X, where the minority class sample set is X_{min} and the majority class sample set is X_{max}.

1. Set the initial cluster center number k of K-means algorithm according to the given data set, randomly select k samples from the minority class samples as the initial cluster center, and use the K-means algorithm to perform the clustering operation on the given data set. Note that when updating the cluster center, the minority class sample point closest to the cluster mean is selected as the new cluster center.
2. In the cluster after the statistical clustering is completed, the number of minority class samples and the number of majority class samples, according to formula (2), formula (3) and formula (4) to determine the category of each cluster (security cluster, noise cluster and boundary cluster).
3. Oversampling the minority class samples of the boundary cluster, taking the cluster center as the core, combining with other original minority class samples in the cluster, and using the interpolation formula (5) to generate new sample points.
4. Undersampling the majority class samples of the boundary cluster, use the Tomek links algorithm to calculate the distance between each minority class sample and all majority class samples, and select the Tomek links sample pair according to formula (6–7), and delete the Tomek links sample pair in the sample pair. majority class sample.
5. Undersampling the majority class samples of the entire dataset, using the NearMiss-2 algorithm to select those majority class samples with the smallest average distance to the three furthest minority samples.
6. Determine whether the data set is close to the balance, if not, iterate the process of steps 3–5 until the data set is close to the balance, the algorithm ends.

4 Experimental Results and Evaluation

In order to verify the effectiveness of the NKB-SMOTE algorithm in network intrusion detection applications, this paper uses the NKB-SMOTE algorithm to generate samples from the UNSW-NB15 dataset, and then trains five common machine learning algorithms for intrusion detection. By comparing the classification accuracy of different sample generation algorithms, the performance of intrusion detection based on NKB-SMOTE algorithm is illustrated. Finally, the actual data acquisition and intrusion detection experiments are carried out for the sewage treatment industrial control system, which illustrates the detection performance of the algorithm in this paper in the real scene.

4.1 Data Preprocessing

The UNSW-NB15 dataset was collected by the Australian Cyber Security Center Laboratory in 2015 through real network environments, recording real modern normal activities and modern integrated attack behaviors [26,27]. Therefore, the UNSW-NB15 dataset can better reflect the characteristics of modern network traffic. The attack types included in this dataset are 9 categories: *Fussers*,

Analysis, *Backdoors*, *Dos*, *Exploits*, *Generic*, *Reconnaissance*, *Shellcode*, and *Worms*. In this dataset, normal traffic data *Normal* accounts for the largest proportion, and attack types *Generic* and *Exploits* account for most of the intrusion data. Therefore, the data distribution of this dataset is unbalanced.

Since the UNSW-NB15 dataset represents traffic records by traffic feature sequences, the representation forms are divided into character type, binary type and numerical type according to different characteristics, and the continuity can be divided into discrete type and continuous type. Machine learning algorithms require data to be only numerical, so before training machine learning classification algorithms, feature data needs to be preprocessed. Preprocessing includes character data digitization and continuous data standardization to ensure data readability and eliminate singular sample data.

1. Numericalization:
   There are some features in the UNSW-NB15 dataset that are useless for intrusion detection, so attributes such as IP address and port number should be removed first, and then each record in the dataset can be equivalent to a 42 feature value and 2 label values. vector. Then, the One-Hot method is used to convert the character data into numerical data, and finally, each record is programmed with 200-dimensional data.
2. Standardization:
   The value ranges of different attributes in the data set are very different, so we need to standardize the data, that is, scale the data to a specific interval first, so that the mean of the data is 0 and the variance is 1.

$$Z_{ik} = \frac{x_{ik} - \bar{x}_k}{S_k}, \tag{8}$$

   where Z_{ik} represents the k-th eigenvalue of the i-th data after z-score normalization, x_{ik} represents the k-th eigenvalue of the i-th data, $\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}$ represents the mean value of the k-th eigenvalue, and $S_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}$ represents the mean absolute deviation of the k-th eigenvalue. Then the data is normalized to the interval $[0, 1]$ by formula (9),

$$Z_{ik}^{'} = \frac{Z_{ik} - Z_{\min}}{Z_{\max} - Z_{\min}}, \tag{9}$$

   where $Z_{ik}^{'}$ represents the k-th eigenvalue of the i-th data after normalization, and $Z_{\min}$ and $Z_{\max}$ represent the minimum and maximum values of the k-th eigenvalue of the data, respectively.

4.2 Experimental Results

On the basis of data preprocessing, this paper carried out a multi-classification experiment of industrial control system intrusion detection. In order to illustrate the effectiveness of the proposed algorithm, five machine learning algorithms were tested on the original dataset and after processing by different sample generation algorithms. Intrusion detection is performed on the data set, as

shown in Table 1, the multi-classification accuracy of five machine learning algorithms under different sample generation algorithms, among which, category 1−9 represents categories *Fussers*, *Analysis*, *Backdoors*, *Dos*, *Exploits*, *Generic*, *Reconnaissance*, *Shellcode*, and *Worms*.

It can be seen that when no sample generation algorithm is used, the five algorithms have poor classification results for some few-category samples. After using the sample generation algorithm, for a few types of attacks, such as 'Analysis', 'Backdoor', 'Shellcode' and 'Worms', the traditional SMOTE algorithm can significantly improve the classification effect of some classification algorithms, such as 'Analysis' attack type classification effect in decision tree algorithm and random forest algorithm, 'Backdoor' and 'Worms' attack type in decision tree algorithm, classification effect of random forest algorithm, etc. The NKB-SMOTE algorithm can improve the classification effect of all classification algorithms for minority class attack samples, and the improvement is more significant and stable. For most types of attack types, such as 'DoS', 'Exploits', 'Generic' attack types, the traditional SMOTE algorithm can not significantly improve its classification effect, and even the classification effect will be slightly worse, such as the classification effect of the 'DoS' attack type in the multilayer perceptron algorithm, the classification effect of the 'Exploits' attack type in the Naive Bayes algorithm, etc. The classification effect of the algorithm, the classification effect of the 'Exploits' attack type in the Naive Bayes algorithm, etc. The NKB-SMOTE algorithm can not only improve the classification effect of the minority class attack samples, but also improve the classification effect of the majority class attack type.

As can be seen from the comparison chart of AUC values in Fig. 6, the average AUC value of the traditional SMOTE algorithm has improved compared with the average AUC value of the original data set, but the improvement is not obvious. In fact, the AUC value of individual attack types has even decreased. The average AUC value of the NKB-SMOTE algorithm has been significantly improved. Compared with the original data set, the average AUC value of the three algorithms has increased by more than 10%, and is basically above 0.9, that is, the detection authenticity is very high.

4.3 Practical Application of NKB-SMOTE Sample Generation Algorithm

In the previous section, the algorithm proposed in this paper is used on the UNSW-NB15 dataset. In order to make the network traffic more realistic, this chapter simulates four common network attacks in the common sewage treatment industrial control system, and captures the real network traffic data. The data is then transformed into a format that can be used for intrusion detection classification through feature extraction. Finally, intrusion detection experiments of different methods are carried out on the new dataset obtained.

The sewage treatment industrial control system has the characteristics of being simple and easy to understand, and is a typical control system in the

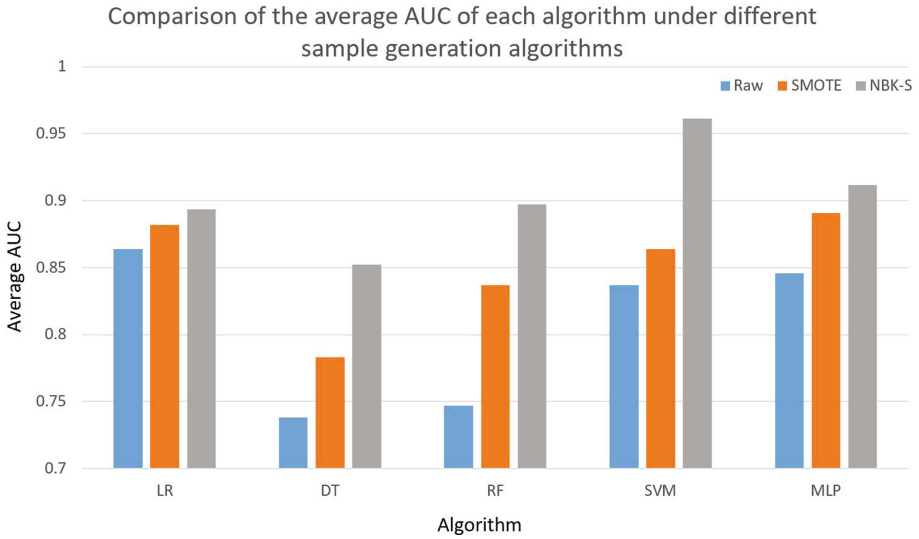


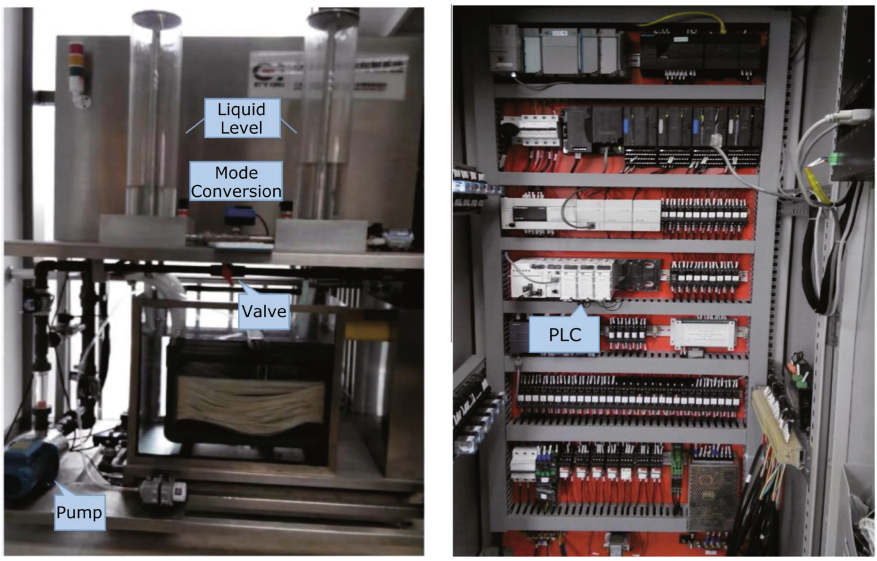
Fig. 6. Comparison of the average AUC of each algorithm under different sample generation algorithms.

process control industry such as water, chemical industry, and petroleum treatment. Therefore, it can be used as a general test platform for information security of industrial control systems. This sewage treatment industrial control system adopts a dual-capacity water tank control mode, which is a typical nonlinear and time-delayed object system. Figure 7(a) and Fig. 7(b) show the equipment diagram of the sewage treatment industrial control system and the control cabinet of the sewage treatment industrial control system, respectively.

By connecting the attack equipment to the sewage treatment industrial control system, this paper simulates different types of attacks to attack the PLC controller, and uses the network packet capture tool Wireshark to sniff and collect network traffic. Then, this paper uses CICFlowMeter to extract features from traffic data and save it as a CSV file. Statistics on data types show that Dos attacks account for the largest proportion, while U2R and R2L attack categories account for a small proportion. Therefore, this dataset is an imbalanced dataset.

Table 1. Multi-classification accuracy of each machine learning algorithm in different sample generation algorithms.

Types	Algorithm	Accuracy (%)				
		LR	DT	RF	SVM	MLP
Fussers	None	93.11	87.28	94.15	98.00	89.67
	SMOTE	89.49	92.03	92.27	96.21	89.70
	NKB-SMOTE	94.13	93.33	94.27	95.38	92.31
Analysis	None	0	10.89	0	0	0
	SMOTE	50.42	67.55	100	26.32	79.54
	NKB-SMOTE	69.27	82.85	90.25	58.38	86.23
Backdoors	None	0	20.60	0	0	0
	SMOTE	43.36	74.33	95.02	54.33	83.02
	NKB-SMOTE	58.92	91.27	84.43	66.30	85.09
DoS	None	33.33	30.33	33.25	33.90	33.79
	SMOTE	34.62	30.87	40.22	57.39	33.25
	NKB-SMOTE	47.62	36.26	46.36	62.84	37.53
Exploits	None	79.00	73.73	69.19	79.83	76.77
	SMOTE	80.31	73.09	66.02	74.27	74.27
	NKB-SMOTE	87.63	81.47	79.62	84.33	80.07
Generic	None	97.02	97.01	99.00	96.34	98.20
	SMOTE	97.41	97.87	99.05	97.37	100
	NKB-SMOTE	97.45	98.03	100	97.78	100
Reconnaissance	None	82.29	89.29	90.11	70.32	76.27
	SMOTE	84.84	83.36	91.27	79.30	83.38
	NKB-SMOTE	88.34	91.25	95.27	83.43	85.02
Shellcode	None	0	42.72	80.20	0	50.04
	SMOTE	19.81	42.63	71.75	38.43	29.62
	NKB-SMOTE	21.98	77.36	84.28	43.73	78.37
Worms	None	0	30.00	50.67	0	0
	SMOTE	55.30	65.47	67.98	74.40	70.42
	NKB-SMOTE	57.17	95.87	95.36	100	88.42



(a) Equipment diagram (b) Control cabinet

Fig. 7. Sewage treatment industrial control system.

Based on the above datasets, we conduct multi-classification experiments on real datasets. Five machine learning algorithms are used to perform intrusion detection on the original data set and the data set processed by different sample generation algorithms. Table 2 shows the multi-classification accuracy of the five machine learning algorithms under different sample generation algorithms. Among them, category 1–4 represents categories DoS, $Probe$, $U2R$, and $R2L$. It can be seen that before and after using the NKB-SMOTE sample generation algorithm in this paper, the intrusion detection accuracy of almost all attack types has been significantly improved, which also shows the effectiveness of the NKB-SMOTE algorithm for real network data collected by the sewage treatment industrial control system, which further proves the effectiveness of intrusion detection based on NKB-SMOTE sample generation in real scenarios.

Table 2. Multi-classification accuracy of each machine learning algorithm in different sample generation algorithms.

Types	Algorithm	Accuracy (%)				
		LR	DT	RF	SVM	MLP
DoS	None	1.00	75.12	95.82	1.00	76.74
	NKB-SMOTE	98.12	93.72	95.47	97.69	93.27
Probe	None	76.67	91.47	92.02	97.04	84.18
	NKB-SMOTE	88.25	95.72	95.73	96.88	91.54
U2R	None	80.96	28.30	31.43	62.53	39.79
	NKB-SMOTE	87.76	90.62	90.47	85.34	87.24
R2L	None	38.13	0	0	42.24	4.18
	NKB-SMOTE	70.40	82.26	84.29	80.37	81.32

In summary, the multi-classification effect of the network data set of wastewater treatment industrial control system has been significantly improved after being sampled by the NKB-SMOTE algorithm, which proves that the NKB-SMOTE algorithm is effective for the real network data collected from wastewater treatment industrial control system, and further proves the effectiveness of the intrusion detection algorithm generated based on NKB-SMOTE sampling in practical scenarios.

5 Conclusion

This paper mainly studies the application of SMOTE sample generation algorithm in intrusion detection. Firstly, the content and defects of traditional SMOTE algorithm are introduced. Then, the NKB-SMOTE algorithm is proposed, and the theoretical basis, research content and algorithm flow of the NKB-SMOTE algorithm are introduced. Finally, in order to verify the application effect of the NKB-SMOTE algorithm in intrusion detection and compare

the advantages of the traditional SMOTE algorithm, based on the UNSW-NB15 dataset, five commonly used machine learning algorithms are used to conduct multi-classification experiments. According to the experimental results, the classification effects of traditional SMOTE algorithm and NKB-SMOTE algorithm for minority class samples and majority class samples in multi-classification scenarios are analyzed respectively. In the future, the application of new machine algorithms in intrusion detection will be considered, and different classification algorithms will be combined in intrusion detection through ensemble learning.

References

1. Index, E.: Global. Nature **522**(7556), S1-27 (2015)
2. Ali, S., Al Balushi, T., Nadir, Z., Hussain, O.K.: Cyber Security for Cyber Physical Systems, vol. 768, pp. 11–33. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75880-0
3. Zhang, J., Pan, L., Han, Q.L., Chen, C., Wen, S., Xiang, Y.: Deep learning based attack detection for cyber-physical system cybersecurity: a survey. IEEE/CAA J. Autom. Sinica **9**(3), 377–391 (2021)
4. Ashfaq, A.B., Javed, M., Khayam, S.A., Radha, H.: An information-theoretic combining method for multi-classifier anomaly detection systems. In: 2010 IEEE International Conference on Communications, pp. 1–5 (2010)
5. Ye, N., Emran, S.M., Chen, Q., Vilbert, S.: Multivariate statistical analysis of audit trails for host-based intrusion detection. IEEE Trans. Comput. **51**(7), 810–820 (2002)
6. Hajji, M., et al.: Multivariate feature extraction based supervised machine learning for fault detection and diagnosis in photovoltaic systems. Eur. J. Control. **59**, 313–321 (2021)
7. Yang, Y., Xu, X., Wang, L., Zhong, W., Yan, C., Qi, L.: Fast anomaly detection based on data stream in network intrusion detection system. In: ACM Turing Award Celebration Conference-China (ACM TURC 2021), pp. 87–91 (2021)
8. Tan, L., Li, C., Xia, J., Cao, J.: Application of self-organizing feature map neural network based on K-means clustering in network intrusion detection. Comput. Mater. Continua **61**(1), 275–288 (2019)
9. Luo, F., Zou, Z., Liu, J., Lin, Z.: Dimensionality reduction and classification of hyperspectral image via multistructure unified discriminative embedding. IEEE Trans. Geosci. Remote Sens. **60**, 1–16 (2021)
10. Kushwaha, P., Buckchash, H., Raman, B.: Anomaly based intrusion detection using filter based feature selection on KDD-CUP 99. In: TENCON 2017–2017 IEEE Region 10 Conference, pp. 839–844 (2017)
11. Jamalipour, A., Murali, S.: A taxonomy of machine learning based intrusion detection systems for the internet of things: a survey. IEEE Internet Things J. **9**, 9444–9466 (2021)
12. Tang, T.A., Mhamdi, L., McLernon, D., Zaidi, S.A.R., Ghogho, M.: Deep learning approach for network intrusion detection in software defined networking. In: 2016 International Conference on Wireless Networks and Mobile Communications (WINCOM), pp. 258–263 (2016)
13. Sun, B., Chen, H., Wang, J., Xie, H.: Evolutionary under-sampling based bagging ensemble method for imbalanced data classification. Front. Comp. Sci. **12**(2), 331–350 (2018). https://doi.org/10.1007/s11704-016-5306-z

14. Lin, W.C., Tsai, C.F., Hu, Y.H., Jhang, J.S.: Clustering-based undersampling in class-imbalanced data. Inf. Sci. **409**, 17–26 (2017)
15. Padmaja, T.M., Krishna, P.R., Bapi, R.S.: Majority filter-based minority prediction (MFMP): an approach for unbalanced datasets. In: TENCON 2008–2008 IEEE Region 10 Conference, pp. 1–6 (2008)
16. Kang, Q., Chen, X., Li, S., Zhou, M.: A noise-filtered under-sampling scheme for imbalanced classification. IEEE Trans. Cybern. **47**(12), 4263–4274 (2016)
17. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
18. Li, J., Zhu, Q., Wu, Q., Fan, Z.: A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors. Inf. Sci. **565**, 438–455 (2021)
19. Dolo, K.M., Mnkandla, E.: Modifying the SMOTE and safe-level SMOTE oversampling method to improve performance. In: Woungang, I., Dhurandher, S.K. (eds.) 4th International Conference on Wireless, Intelligent and Distributed Environment for Communication. LNDECT, vol. 94, pp. 47–59. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-89776-5_4
20. He, H., Bai, Y., Garcia, E. A., Li, S.: ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322–1328 (2008)
21. Han, H., Wang, W.-Y., Mao, B.-H.: Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: Huang, D.-S., Zhang, X.-P., Huang, G.-B. (eds.) ICIC 2005. LNCS, vol. 3644, pp. 878–887. Springer, Heidelberg (2005). https://doi.org/10.1007/11538059_91
22. Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J.: Hybrid sampling for imbalanced data. Integr. Comput.-Aided Eng. **16**(3), 193–210 (2009)
23. Li, H., Zou, P., Wang, X., Xia, R.: A new combination sampling method for imbalanced data. In: Sun, Z., Deng, Z. (eds.) Proceedings of 2013 Chinese Intelligent Automation Conference. LNEE, vol. 256, pp. 547–554. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38466-0_61
24. Dina, A.S., Manivannan, D.: Intrusion detection based on Machine Learning techniques in computer networks. Internet Things **16**, 100462 (2021)
25. Ahmad, Z., Shahid Khan, A., Wai Shiang, C., Abdullah, J., Ahmad, F.: Network intrusion detection system: a systematic study of machine learning and deep learning approaches. Trans. Emerg. Telecommun. Technol. **32**(1), e4150 (2021)
26. Moustafa, N., Slay, J.: The evaluation of network anomaly detection systems: statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set. Inf. Secur. J. Glob. Perspect. **25**(1–3), 18–31 (2016)
27. Moustafa, N., Slay, J.: UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In: 2015 Military Communications and Information Systems Conference (MilCIS), pp. 1–6 (2015)