Deep Learning Intrusion Detection Model Based on Optimized Imbalanced Network Data

Yan Zhang, Hongmei Zhang, Xiangli Zhang, Dongsheng Qi
School of Information and Communication
Guilin University of Electronic Technology
Guilin, China
e-mail: 1512898308@qq.com, Hmzhang@guet.edu.cn, xlzhang@guet.edu.cn, dongsheng_q@163.com

***Abstract*—To solve the problem of the low detection rate of minority samples in imbalanced datasets in network intrusion detection, a deep learning intrusion detection model based on optimized imbalanced data is proposed. Firstly, a hybrid sampling method is adopted in data processing. Synthetic Minority Over-sampling Technique (SMOTE) was used to increase the numbers of samples in minority categories and the majority of the samples were under-sampled by Neighborhood Cleaning Rule (NCL). Secondly, on the preprocessed balanced dataset, the high-dimensional data was reduced by Deep Belief Network (DBN) to obtain the lower low-dimensional representation of the preprocessed data. Finally, the classification work was completed by Probabilistic Neural Network (PNN). The experiment on NSL-KDD dataset showed that hybrid sampling can improve the detection rate and classification accuracy of minority categories. And the performance of DBN-PNN is obviously superior to the traditional method.**

Keywords-intrusion detection; Synthetic Minority Over-sampling Technique; Neighborhood Cleaning Rule; Deep Belief Network; Probabilistic Neural Network.

I. INTRODUCTION

In recent years, with the continuous expansion of network scale and the gradual updating of network technology, the information leakage events at home and abroad have occurred repeatedly. So Intrusion detection has become an inevitable key issue.

Many researchers have introduced a variety of machine learning methods for this problem, such as neural network (NN) [1], support vector machine (SVM) [2] and so on, which have made great achievements. However, in the face of massive complex data, traditional machine learning methods have limited expression ability, and feature learning is easily constrained by time and space complexity, resulting in low accuracy and high false alarm rate.

Deep learning has a prominent performance in massive data analysis, which can be used to solve the intrusion detection problems in complex network environment. In 2006, Hinton first proposed the deep belief network (DBN) [3], which made a breakthrough in image classification and speech recognition [4]. In the last few years, some researchers have applied DBN to intrusion detection. Gao et al. proposed DBN as a classification model for intrusion detection for the first time and verified that DBN can improve the detection effect [5]. Alom et al. applied the DBN model only for feature dimensionality reduction and used SVM algorithm to classify the data after dimensionality reduction, the classification accuracy was higher than using the single DBN or SVM [6]. However, the work of network intrusion detection is mostly based on the datasets of KDD99 or NSL-KDD, both of which are serious imbalance. Most deep learning methods aim to improve the overall detection rate without considering the detection problem of minority samples, resulting in low detection rate of minority categories.

In view of the above problems, this paper proposed a hybrid deep learning intrusion detection model based on optimized unbalanced data. Firstly, in the data processing stage, in response to balance the sample data of intrusion detection, SMOTE and NCL is used to form a balanced dataset. Then, DBN is used to reduce the dimension of the features on the balanced dataset and retain the key features of the data. Finally, the probability neural network (PNN) is appropriate to classify the low dimensional data. DBN-PNN model can achieve better results when dealing with massive and complex data. In this paper, experiments are carried out on the NSL-KDD dataset results showed that mixed sampling can significantly improve the detection accuracy of minority samples, and the model has higher detection accuracy, which proves that the model proposed has better performance.

II. CORRELATION ALGORITHM

A. SMOTE

The SMOTE algorithm proposed by Chawla is one of the classical over-sampling techniques [7]. The main idea of this method is to increase the number of minority class samples by inserting a small number of synthetic samples into the samples which are close to the minority class samples, and it can solve the problem of over-fitting classification effectively.

S is a set of samples and X is a small number of samples. $X \subset S$, $x_i \subset X$ .First, the over-sampling ratio N is set to find the nearest neighbors of the same class k for each x_i . Then, n new samples are synthesized according to

equation (1). Finally, the new samples generated by the algorithm are added to the set S .

$$x_{new} = x_i + rand(0,1) * (y_j - x_i) \quad (1)$$

Among equation (1), $j = 1,2,\cdots n$, x_{new} represents samples after over-sampling, y_j represents the nearest neighbor of x_i which are k of them, $rand(0,1)$ represents a random number in region (0,1).

B. NCL

In the case that the traditional random under-sampling does not consider the distribution of samples, most of the important sample information may be deleted. Laurikkala proposed neighborhood cleansing (NCL) algorithm [8].

Main steps:

Step1: For element x_i in dataset S , k nearest neighbor K_i of x_i is calculated.

Step2: For element K_i in k_i , if k_i is different from x_i class, count $sum = sum + 1$ and add k_i to D .

Step3: If x_i is a minority group and $sum \geq k - 1$, deleted D from S and that is $S' = S - D$.

Step4: If x_i is a majority group and $sum \geq k - 1$, deleted x_i from S and that is $S^* = S - x_i$.

Because of the huge amount of data in the intrusion detection dataset, it is slow to under-sampling the majority of the class samples by NCL. So the original dataset S is divided into n parts $S_1, S_2, \cdots S_n$ randomly by grouping. The majority of class samples in each dataset are under-sampling by NCL algorithm. Grouping can accelerate under-sampling speed significantly.

C. Deep Belief Network

DBN is a deep neural network composed of a multi-layer unsupervised Restricted Boltzmann Machine (RBM) network and a supervised BP neural network. In the process of training model, DBN is divided into two steps, namely, pre-training and fine-tuning. RBM is the basic component of DBN. It is a generative stochastic neural network with two layers. The first layer is called the visual layer, and the second layer is the hidden layer. There are connections between all the visible and hidden layers. While the visual layer units are not connected with each other .The RBM structure is shown in Fig.1, v and h represent visible layer and hidden layer respectively, a and b are the offsets of visible layer and hidden layer, w represents the connection weight between the two layers. All units are two valued variables.

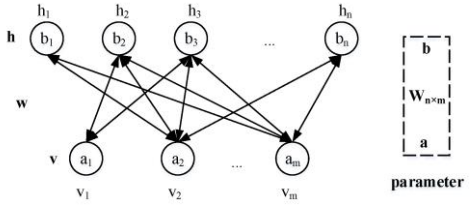

Figure 1. RBM structure

The m and n are the number of nodes in the visible and hidden layers, v_i represents the state value of the i-th visible layer unit, h_j represents the state value of the j-th hidden layer unit. RBM is an energy based model whose energy function can be expressed as:

$$E(v,h \mid \theta) = -\sum_{i=1}^{m} a_i v_i - \sum_{j=1}^{n} b_j h_j - \sum_{i=1}^{m}\sum_{j=1}^{n} v_i h_j w_{ij} \quad (2)$$

where $\theta = (w_{ij}, a_i, b_j)$ is the model parameter of RBM. w_{ij} represents the connection weight between the i-th unit visible layer and the j-th hidden layer unit weight, a_i represents the offset of the i-th visible layer unit, b_j represents the offset of the j-th hidden layer unit.

Comparison divergence (CD) algorithm is a fast training RBM method [9]. After the complete algorithm, the updated criterion of the parameters is obtained.

$$\begin{cases} w_{ij} = \varepsilon(\langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{recon}) \\ a_i = \varepsilon(\langle v_i \rangle_{data} - \langle v_i \rangle_{recon}) \\ b_j = \varepsilon(\langle h_j \rangle_{data} - \langle h_j \rangle_{recon}) \end{cases} \quad (3)$$

ε is the learning rate in the training process, $\langle \cdot \rangle_{data}$ is the mathematical expectation on the distribution defined for the training dataset, $\langle \cdot \rangle_{recon}$ is the expectation of the distribution defined in the reconstructed model.

In the DBN model of this paper, BP network plays a fine role in adjusting weights. The specific process is as follows: for each training sample x_i , input DBN network, calculate the actual output value o_i of x_i , and the expected output value is t_i , then its error δ is:

$$\delta_i = o_i(1 - o_i)(t_i - o_i) \quad (4)$$

For the s hidden layer of DBN, the actual output of the i node is y_i , and the error term δ^s is calculated.

$$\delta_i^s = y_i^s(1 - y_i^s)\sum_j w_{ij}^l \delta_j^{s+1} \quad (5)$$

At this point, you can update the network parameters of DBN as follows, ε is the learning rate of the fine-tuning stage.

$$\begin{cases} w_{ij}^s = w_{ij}^s + \varepsilon_{fine-tuning} \times y_i^s \delta_j^{s+1} \\ b_j^s = b_j^s + \varepsilon_{fine-tuning} \times \delta_j^{s+1} \end{cases} \quad (6)$$

D. PNN

PNN is a feedforward neural network based on Bayesian minimum criterion [10]. It is simple in structure, easy in training, fault-tolerant and widely used. In practical applications, especially in solving classification problems, the advantage of PNN network is that it uses linear learning algorithm to complete the work of nonlinear learning algorithm, and can maintain the high accuracy of nonlinear algorithm. The PNN network is divided into four layers: input layer, mode layer, summation layer and output layer. Its structure is shown in Fig. 2.

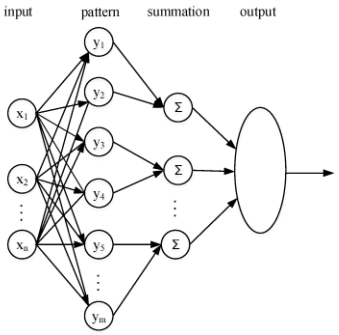

Figure 2. PNN structure

III. INTRUSION DETECTION MODEL

A. Model Design

To sum up, the model framework presented in this paper is illustrated in Fig. 3, which contains 3 steps.

1) Data preprocessing:

- Firstly, the symbolic data in NSL-KDD dataset is transformed into numeric data, and all numeric data are normalized.
- Aiming at a few kinds of samples in the dataset after the preprocessing, the number of them is increased by SMOTE, and the majority of the samples in the data set are under-sampled by NCL. After sampling, a balanced data set is formed.

2) Data dimensionality reduction: The preprocessed data will be dimensionality reduced by DBN model. The specific process is as follows: firstly, unsupervised CD algorithm is used to train each RBM from bottom to top layer to remove the redundant features of the original data, obtain the initial parameters of DBN model and low-dimensional data, and then fine-tune the whole DBN network from top to bottom by BP algorithm to get the optimal model parameters.

3) Data classification: use PNN network to classify low dimensional output data.

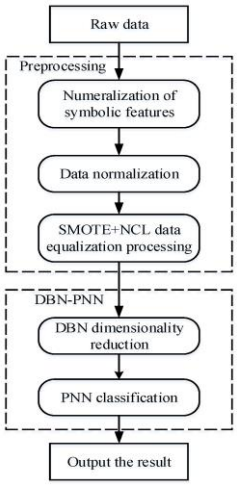

Figure 3. DBN-PNN intrusion detection framework based on optimized imbalanced data

B. Data Preprocessing

1) Numeralization of symbolic features

Each record in the NSL-KDD dataset contains 41 attribute features and one attribute label identifying the attack category. The 41 attributes are divided into 38 numeric features and 3 symbolic features. Symbolic properties are transformed into numeric data by attribute mapping. For example, attribute feature "protocol_type" has three values: tcp,udp,icmp, encoding as [1,0,0], [0,1,0] and [0,0,1]. Similarly, attribute features "service" and "flag" can establish a mapping relationship between symbolic values and corresponding values. Finally, the feature changes from 41 to 122 dimensions by mapping.

2) Normalization of numeric features

In order to eliminate the effect of dimension for every single attribute, the training set and the testing set must be normalized, and the data size is normalized to the range of [0,1] according to equation [7]:

$$y = \frac{y - MIN}{MAX - MIN} \tag{7}$$

where y is a numerical value, MIN is the minimum value for the attribute that y belongs to, and MAX is the maximum value for the attribute that y belongs to.

3) Dataset equalization processing

As shown in Table I, in the original training set, the proportion of sample categories is seriously unbalanced, and the number of Normal, Dos and Probe is much larger than the number of U2R. Therefore, a mixed sampling method is adopted for data sets, that is, SMOTE oversampling is applied to a small number of samples U2R to increase the data volume by 20 times, and NCL under-sampling is carried out for most samples (Normal, Dos and Probe). In under-sampling, because the total data set is huge, the grouping method is adopted. The over-sampled data is divided into n copies, where n is taken 10, because the information loss caused by under-sampling can be minimized by dividing into 10 copies, and a balanced data set is finally obtained on the basis of guaranteeing the original data distribution characteristics. The hybrid sampling algorithm is designed as follows:

Algorithm 1: hybrid sampling algorithm

Input: The original non-equilibrium dataset S, the over-sampling rate N;

Output: Balanced dataset M.

Step1: According to the over-sampling rate N, the SMOTE algorithm (see the section 1.1) is used to enlarge the minority samples and get the dataset S'.

Step2: S' is randomly divided into 10 parts: $S'_1, S'_2, \cdots S'_{10}$. First, the Normal samples of each $S'_k (1 \le k \le 10)$ are sampled by NCL (see the section 1.2) and were under-sampling by NCL and $O'_1, O'_2 \cdots O'_{10}$ were obtained.

Step3: Similarly, Dos and Probe in O'_k are under-sampled respectively, and finally get data $P_1, P_2, \cdots P_{10}$.

Step4: The resulting data $P_1, P_2, \cdots P_{10}$ is aggregated into a table, which is to balance the data set M and output M .

TABLE I. NSL-KDD DATASET DISTRIBUTION

Data Types	Training data	Testing data
Normal	67343	9711
Probe	11656	2421
Dos	45927	7458
U2R	52	200
R2L	995	2754

IV. EXPERIMENT AND RESULT ANALYSIS

A. Dataset Description

This paper uses the NSL-KDD dataset [11], which is an improved version of KDD99. It removes some redundant data on the basis of KDD99, and is more suitable for intrusion detection experiments. The NSL-KDD dataset contains 125973 training data and 22544 test data, including one normal class of data and four types of attack data: Dos, R2L, U2R and Probe.

B. Setting of Experimental Parameters

In the process of experiment, when the DBN model is trained on the data, different parameter settings of DBN will affect the training result of the model, and then influence the classification effect. In this paper, according to the reference [12] and the experimental results of many tests, the DBN model and other parameters are set as shown in Table II.

TABLE II. EXPERIMENTAL PARAMETERS

DBN Model Parameters	Values
input layer nodes	122
the first hidden layer nodes	90
the second hidden layer nodes	60
output layer nodes	30
learning rate of pre-training	0.05
learning rate of fine-tuning	0.5
iterations of pre-training	30
iterations of fine-tuning	150

C. Evaluation Measurement

Generally, accuracy (AC), detection rate (DR) and false alarm rate (FA) are used in the intrusion detection field as indicators to measure the performance of intrusion detection model:

Accuracy: the proportion of the data that is correctly classified as the total test data; Detection rate: the proportion of the intrusion data that is correctly classified as the total invasive data; False alarm rate: the proportion of data that is misclassified in normal data to total normal data.

However, this traditional evaluation standard is not accurate for unbalanced data. Because classifiers are insensitive to minority classes, the recognition rate of minority classes is very low when classifying unbalanced data. So there are some new evaluation indexes for unbalanced data: AUC, F-value and G-means. The AUC index was used in this experiment. AUC is the area below the corresponding ROC curve, which can take into account the detection rate of most classes and minority classes. The greater the value of AUC, the higher the accuracy of classification, the better the effect.

TP: the number of samples correctly judged as positive samples.

TN: the number of samples correctly judged as negative samples.

FN: the sample error is judged as the actual negative sample number of positive class.

FP: the number of sample positive samples is negative.

The AUC is defined as follows:

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2} \tag{8}$$

$$FP_{rate} = FP / (FP + TN) \tag{9}$$

$$TP_{rate} = TP / (TP + FN) \tag{10}$$

D. Experimental Analysis

1) Experimental Environment: Windows10 (64bits) operating system, Matlab R2017a.

2) Experimental Content

This paper designs two groups of experiments: verify the influence of mixed sampling algorithm on Intrusion Detection Based on the same classification method; compare the intrusion detection model and other classification methods on the basis of balanced dataset.

a) The influence of mixed sampling algorithm on intrusion detection

This experiment uses three methods of preprocessing data to carry on the contrast experiment. The specific three types of data are: the raw data that has not been processed, and recorded as untreated; the data processed by the SMOTE algorithm is denoted as SMOTE; the data processed by mixed sampling algorithm is denoted as mixed sampling. In order to ensure the same amount of data in the experiment, the data of the three sampling methods are 20% of the original training set samples, and the balanced dataset distribution formed by the mixed sampling after SMOTE over-sampling and NCL under sampling is shown in Table III. The experiment uses three kinds of data as training data to train the model, and finally analyzes the performance on the testing set. Table IV compares the ACU values obtained by experiments on testing set using untreated, SMOTE, and mixed sampling algorithms.

TABLE III. DATASET AFTER MIXED SAMPLING

Data Types	Normal	Probe	Dos	U2R	R2L
Number	9477	5410	8026	1092	995

TABLE IV. COMPARISON OF EXPERIMENTAL RESULTS OF VARIOUS SAMPLING METHODS

Sampling method	AUC (Normal)	AUC (Probe)	AUC (Dos)	AUC (U2R)	AUC (R2L)
Untreated	0.9768	0.9798	0.9950	0.7934	0.9351
SMOTE	0.9665	0.9798	0.9886	0.9427	0.928
Mixed sampling	0.9837	0.9866	0.9953	0.9452	0.9467

According to the results in Table IV, compared with the unprocessed data and the SMOTE algorithm, the AUC values of all kinds of mixed samples are the highest. The AUC value of the mixed sampling U2R is about 15% higher than that of the unprocessed data, and is only 0.25% higher than that of the SMOTE method, but overall, the AUC of all kinds of mixed sampling is higher than that of the other two methods. Therefore, the intrusion detection model based on mixed sampling has better performance in dealing with the classification results of unbalanced data.

b) Performance comparison with other classification models

In order to prove that DBN-PNN has better performance than original DBN and PNN, experiments were carried out to explore the effects of training each model on the balanced dataset after mixed sampling, in which DBN and DBN-PNN adopt the same parameters.

TABLE V. EXPERIMENTAL RESULT

Model	AC	DR	FR
DBN	96.48%	96.35%	3.33%
PNN	97.94%	98.05%	3%
DBN-SVM	98%	99.05%	3.54%
DBN-PNN	98.32%	99.35%	3.2%

Table V shows that the accuracy and detection rate of DBN-PNN model proposed in this paper are higher than those of DBN, PNN and DBN-SVM methods, and the false alarm rate is only slightly higher than that of PNN method, but lower than other methods. Considering the various indicators, the performance of this model is better and more stable.

V. CONCLUSION

In this paper, a deep learning intrusion detection model based on optimized unbalanced data is proposed, which improves the detection ability of minority classes in intrusion detection. Firstly, the minority samples are increased by SMOTE technology, and then under-sampling the majority of samples by NCL, so as to get a balanced dataset, which solves the problem of low detection rate of minority categories. At the same time, the DBN model is improved. Combining the advantages of strong classification ability, accuracy and simple training, the DBN-PNN model is proposed. Experimental results on balanced dataset show that the proposed method has better performance than the traditional methods. Because the structure and network parameters of DBN have many uncertainties, the detection rate will be affected by the learning rate, the number of iterations and other factors, so the next step to solve the problem is how to effectively select model parameters.

ACKNOWLEDGMENT

This work was supported by the Foundation items (NFS of China: No. 61461010, No.61363031); Key Laboratory of Cognitive Radio and Information Processing, Ministry of Education (Guilin University of Electronic Technology) (No.CRKL170103, No. CRKL170104).

REFERENCES

[1] B. Shah, H. B. Trivedi, "Artificial Neural Network based Intrusion Detection System: A Survey," International Journal of Computer Applications, Vol. 39, no. 6, p.p. 13-18, 2012.

[2] F. Kuang, W. Xu, S. Zhang, "A novel hybrid KPCA and SVM with GA model for intrusion detection," Applied Soft Computing, Vol. 18, p.p. 178-184, 2014.

[3] G. E. Hinton, S. Osindero, Y.-W. Teh, "A fast learning algorithm for deep belief nets," Neural Computation, Vol. 18, p.p. 1527-1554, 2006.

[4] D. Yu, L Deng, "Deep Learning and Its Applications to Signal and Information Processing [Exploratory DSP]," IEEE Signal Processing Magazine, Vol. 28, no. 1, p.p. 145-154, 2010.

[5] N. Gao, L. Gao, Q. Gao, et al, "An Intrusion Detection Model Based on Deep Belief Networks," Second International Conference on Advanced Cloud and Big Data. IEEE Computer Society, p.p. 247-252, 2014.

[6] M. Z. Alom, V. R. Bontupalli, T. M. Taha, "Intrusion detection using deep belief networks," Aerospace and Electronics Conference. IEEE, p.p. 339-344, 2016.

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, et al, "SMOTE: synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, Vol. 16, no. 1, p.p. 321-357, 2002.

[8] J. Laurikkala, "Improving Identification of Difficult Small Classes by Balancing Class Distribution," Conference on Ai in Medicine in Europe: Artificial Intelligence Medicine. Springer-Verlag, p.p. 63-66, 2001.

[9] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," MIT Press, 2002.

[10] P. O. Chasset, "pnn: Probabilistic neural networks," MIT Press, p.p. 109-118, 2013.

[11] L. Dhanabal, S. P. Shantharajah, "A study on NSL-KDD dataset for intrusion detection system based on classification algorithms," International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, no. 6, p.p. 446-452, 2015.

[12] N. Gao, Y. He, L. Gao, et al, "Deep learning method for intrusion detection in massive data," Application Research of Computers, 2018.