ISSN:2790-1688 Volume-15-(2025)

A Neural Network Fusion Feature Selection for Network Intrusion Detection

Wenhao Jiang[1,a], Zheng Li[2.b]

[1]College of Intelligent Information Engineering, Chongqing Aerospace Polytechnic, Chongqing, China

[2]Mobile Communication Engineering Research Center, Chongqing University of Posts and Telecommunications, Chongqing, China

[a]yljdp@yeah.net, [b]lizheng@cqupt.edu.cn

Abstract. Network Intrusion Detection often utilized to shield computer networks from a multitude of attacks, poses a significant challenge in the realm of system and network security due to the pervasive nature of network threats. At present, Deep Learning methods of Neural Networks are often used in Network Intrusion Detection, but the performance of classifiers may vary from data set to data set. The main reason for this is that there are some errors or duplicate features. To solve this problem, three feature selection methods are combined with Neural Networks to improve the performance of Network Intrusion Detection. In the experimental evaluation phase, this paper uses experimental results demonstrate the superiority of the proposed method in comparison to existing Network Intrusion Detection techniques, by identifying significant and intimately related features as tested on the KDD99 dataset for feasibility.

Keywords: Neural Network, Intrusion Detection, Feature Selection.

1. Introduction

The normal operation of network systems is seriously hindered by a multitude of threats that network security encounters, including denial of service attacks, computer viruses and data leaks. Widely implemented to counteract these assaults, the Network Intrusion Detection System (IDS) deploys varying detection technologies, network intrusion detection system categorizes into two types, which are the anomaly-based Network Intrusion Detection System (NIDS) and the feature-based Network Intrusion Detection System. Network Intrusion Detection based on anomaly is capable of uncovering unknown attacks within the network, while feature based Network Intrusion Detection is mainly used to detect known network attacks. Machine Learning plays an important role in anomaly based Network Intrusion Detection. Machine Learning detects potential network attacks by training classifiers using historical data to create profiles that recognize anomalies. Neural Network models including the Radial Basis Function Network(RBFN) and Back Propagation Neural Network(BPNN) are widely used in anomaly-based Network Intrusion Detection Systems. However, due to the high dimensionality and instability of the dataset, the performance of Machine Learning classifiers can be affected to varying degrees. Redundant or irrelevant features diminish detection performance and effective feature selection can address this issue, particularly in support vector machine classifiers.

1.1 Main Contributions

The main contributions of this paper are as follows:

Three main feature selection techniques: relevance, fast correlation and consistency, are proposed to select useful features and reduce data redundancy, improve the efficiency and effectiveness of feature selection, and thus improve the overall performance of Network Intrusion Detection.

RBFN and BPNN are integrated with feature selection to create an efficient method for Network Intrusion Detection, enhancing detection accuracy and significantly reducing computing costs.

Conduct evaluation on the KDD99 dataset to verify the practicability and effectiveness of the proposed method. The experimental results show that the method has good detection performance on different data sets, and the method has wide applicability.

1.2 Article Structure

The rest of the article is structured as follows:

The second part provides a comprehensive review of relevant research, introduces Machine Learning methods, and focuses on the application of Neural Networks in the field of Network Intrusion Detection.

The third part puts forward the method of feature selection and neural network fusion, and elaborates the network intrusion detection process in detail.

The fourth part presents the experimental process and results, evaluating the performance of the proposed method and comparing its predictive capabilities.

The fifth part summarizes the research methods used in this paper, evaluates the application of bioanalysis, and discusses future research directions and potential improvement measures.

2. Related work

In Network IDS, Machine Learning plays a crucial role in identifying intrusions. Researchers employ classification techniques, including Decision Trees (DT), Neural Networks (NN), and Support Vector Machines (SVM), to enhance the performance of Network Intrusion Detection. The following is a review of the relevant research.

2.1 Neural Network

In the field of network intrusion detection, neural network has shown its strong ability. P. Kanagavalli and V. Karthikeyani[1] proposed a correlation-based feature extraction (CFS) algorithm, which is an innovative method for extracting key features from network data. They further combined this algorithm with the Trust Algorithm (TA) to assess the trustworthiness of network nodes. In addition, they used the secure Random Forest algorithm (SRFA) as a classifier to classify nodes into three categories: trusted, untrusted, or malicious. Experimental data show that this intrusion detection system can effectively reduce the false positive rate, showing its superiority in identifying network anomalies, surpassing other existing systems.

On the other hand, Ahmad et al.[2] developed an Adaboost-based network intrusion detection technique that focuses on feature selection and achieves a high accuracy of 99.3% on the UNSW-NB15 dataset. The system monitors network traffic and classifies it into threat and non-threat categories, which provides strong support for the application and research of network security.

Wang et al. [3] used short term memory network (LSTM) to process time series data and introduced a layer-by-layer training strategy to improve the performance and stability of the model. Their experimental results perform well on the KDD99 dataset, further confirming the effectiveness of this method in network intrusion detection.

2.2 Decision Tree

Decision Tree algorithm is widely used in the research of network intrusion detection system (NIDS). N Kumar and U Kumar[4] successfully improved the detection efficiency of NIDS by integrating multiple techniques such as feature selection, data normalization, fuzzy C-means clustering and C4.5 decision tree. Their experiments on the KDD99 dataset show that this comprehensive approach can achieve excellent detection results.

At the same time, VG Krishnan et al.[5] investigated the application potential of deep learning techniques in NIDS, in particular combining one-dimensional convolutional neural networks(1D-CNN) with Chimpanzee optimization algorithm(COA) for feature extraction, and constructed a hierarchical network model (HNM). The experimental results on the NSL-KDD database confirmed the superiority of the CNN-COA model in classification accuracy, and showed better performance compared with the prior art. The model achieved 87.19% accuracy and 88% to 89% accuracy and recall, which represents a significant improvement over the existing CNN model with 81.75% accuracy and 82% accuracy and recall.

2.3 Support Vector Machine

In the research field of network intrusion detection system (NIDS), support vector machine (SVM) plays a key role. Zolbayar et al.[6] designed an attack algorithm called NIDSGAN based on Generative Adversarial Networks (GANs) to evaluate the effectiveness of Machine Learning based NIDS in the real world. Their research faces two core challenges. First, network features need to conform to domain constraints to ensure that they can truly reflect network behavior. Second, the attacker needs to learn and simulate the decision-making process of the target NIDS model without knowing the internal details, such as the architecture and parameters. Through this work, Zolbayar et al. provided a new perspective and method for the security assessment of NIDS.

3. Proposed Method

In this section, we introduce the Network Intrusion Detection method that combines Neural Networks with feature selection. The main components of this approach include data preprocessing, a selection of three feature selection techniques, and a fused Neural Network model.

3.1 Data Preprocessing

Data preprocessing is the key step of subsequent data analysis, and the realization of data preprocessing includes the following three steps.

Data cleaning deals with missing values and outliers. Means of filling, interpolating, or deleting relevant records for missing data. Statistical methods or rules based on domain knowledge are used to deal with outliers.

Data normalization scales the data to a unified range to eliminate the impact of dimensional differences on model training. The common methods of data normalization include min-max normalization and Z-score standardization.

Data segmentation which are the data set is divided into training set, verification set and test set. 70% of the data in the data set is used for model training, 15% for model verification and 15% for model testing. Data sets are usually split in such proportions to ensure the generalization ability of the model.

3.2 Neural Network Model

In this paper, the RBFN which performs complex mapping, is used to classify the data by calculating the similarity between the input sample and the training set sample. Each neuron in the training set is retained as the base instance, and the Euclidean distance between the input and the base instance is calculated when there is a new input. The advantage of BPNN is the ability to incorporate differentiable transfer functions at each network node, adjust the internal network weights by error backpropagation at the end of each training cycle, and modify the network weights using the backpropagation of BPNN Neural Networks.

3.3 Data Feature Selection

Data is the cornerstone of building machine learning models and training classifiers. The ideal data is high quality data, but there are many challenges with data in the real world, for example, the "curse of dimension" problem refers to the fact that the number of regions in space grows exponentially as the number of dimensions increases, but the data set is always limited. In this paper, fast correlation filtering (FCBF), correlation-based filtering (CBF) and consistency method (CM) are used for data feature selection.

3.3.1 Fast Correlation Filtering (FCBF)

The two main goals of FCBF are to reduce redundancy and increase the relevance of input features to classes. Specify $S_i = F - F_i$ as a subset of the set of F features missing F_i, and F_i as a specific feature.

Fast correlation can be divided into C correlation and F correlation. The correlation between categories and features is called C-correlation, and the correlation between a set of features is called F-correlation.

C correlation solution is shown in equation (1).

$$C-correlation: SU_{i,Y} = SU(F_i, Y) \tag{1}$$

F correlation solution is shown in equation (2).

$$F-correlation: SU_{i,j} = SU(F_i, F_j) \tag{2}$$

The construction of an approximate Markov chain is shown in equation (3).

$$SU_{j,Y} \geq SU_{i,Y} \wedge SU_{i,j} \geq SU_{i,Y} \tag{3}$$

The correlation solution is shown in equation (4).

$$SU_{j,Y} \geq \gamma \wedge P(Y \mid F_i, S_i) \neq P(Y \mid S_i) \tag{4}$$

The relevant feature is the dominant feature if and only if the relevant feature does not have any approximate Markov chains in the current feature set.

The FCBF algorithm consists of three steps, the first step selects a major feature, the second step removes all features from the approximate Markov chain, and the third step repeats the steps of the first and second steps above until only the dominant feature is in the input data.

3.3.2 Correlation-based Filtering (CBF)

CBF is used to reduce associations between feature sets and category sets.

The evaluation feature subset is shown in equation (5).

$$M_S = \frac{n\overline{r}_{Yf}}{\sqrt{n + n(n-1)\overline{r}_{ff}}} \tag{5}$$

Where M_s is the fraction of the feature subset S_n, n is the number of features in S, $\overline{r}_{Yf}$ is the average correlation between the class set Y and the features in S, $\overline{r}_{ff}$ is the average correlation in S. The best first search strategy is used to determine the optimal subset of all features, and if five consecutive fully extended subsets do not improve, the operation is terminated.

3.3.3 Consistency Method (CM)

Consistency in statistics means that as the population sample size grows, the estimate of the population distribution may converge to the actual population mean. The concept of consistency can also be used to identify subsets of data features.

The definition of consistency is shown in equation (6).

$$\text{Consistency} = 1 - \text{Discrepancy} \tag{6}$$

The definition of inconsistency is shown in equation (7).

$$\text{Discrepancy} = \frac{\text{number of inconsistent examples}}{\text{total number of examples}} \tag{7}$$

Monotonicity is the advantage of a consistent approach. Suppose there are subsets {K0,K1,...,Kn}, the measure of consistency expressed by U is shown in equation (8).

$$\text{if} \quad K_0 \subset K_1 \subset \text{L} \subset K_n \Rightarrow U(K_0) \leq U(K_1) \leq \text{L} \leq U(K_n) \tag{8}$$

This provides the opportunity to apply the best first search, where m is the cardinality of the feature set, to determine the best subset of 2m feature subsets.

4. Evaluation method

In this section, KDD99 data set is used to verify the performance of the proposed neural network fusion feature selection Network Intrusion Detection method.

4.1 KDD99 Data Set

The KDD Cup 1999 dataset (KDD99) is a dataset extracted from packet tracks in military network systems, and despite being more than 10 years old, this network intrusion detection dataset

remains one of the most popular. Therefore, KDD99 was chosen to evaluate the performance of the proposed method. The KDD99 entire dataset had 4,568,533 records, each with 41 characteristics, and the system was evaluated using a randomly selected collection of 39,066 records, each containing 41 characteristics as well as denial of service, R2L, U2R, and detection of four different types of attacks. The 41 features of KDD99 data set include 4 categories as shown in table 1.

Table 1 Features and categories in dataset KDD99

Category Specific	TCP connection basic characteristic	TCP connection content characteristic	Time-based network traffic	Host-based network traffic
Features	duration, protocol_type, service, flag,src_bytes, dst_bytes, land, wrong_fragment, urgent	hot, num_failed_logins, logged_in, num_compromised, root_shell, su_attempted, num_root, num_file_creations, num_shells, num_access_files, num_outbound_cmds, is_hot_login, is_guest_login	count, srv_count, serror_rate, srv_seror_rate, rerror_rate, srv_rerror_rate, same_srv_rate, diff_srv_rate, srv_diff_host_rate	dst_host_count, dst_host_srv_count, dst_host_same_srv_rate, dst_host_diff_srv_rate, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate, dst_host_serror_rate, dst_host_srv_serror_rate, dst_host_srv_serror_rate, dst_host_rerror_rate, dst_host_srv_rerror_rate

4.2 Measurement Index

Metrics are an important tool for evaluating the performance of a model and can help derive these metrics. We can use a number of performance metrics to determine which is better. Here are some basic and important definitions that can help derive these metrics.

TP(True Positive) : the number of positive classes correctly predicted to be positive.

TN(True Negative) : the number of negative classes correctly predicted to be negative.

FP(False Positive) : the number of negative classes that are incorrectly predicted to be positive.

FN(False Negative) : the number of positive classes that are incorrectly predicted to be negative.

Accuracy Rate : accuracy rate refers to the proportion of the number of correctly predicted samples to the total number of samples, indicating the overall prediction ability of the model. The accuracy rate formula is shown in equation (9).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

Recall Rate : recall rate measures the model's ability to identify positive samples and represents the proportion of all actual positive samples that are correctly identified. The recall rate formula is shown in equation (10).

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{10}$$

False Positive Rate : false positive rate reflects the proportion of negative classes incorrectly predicted by the model to be positive, representing the proportion of all actual negative classes incorrectly predicted to be positive. The false positive rate formula is shown in equation (11).

$$FPR = \frac{FP}{FP + TN} \tag{11}$$

False Negative Rate : false negative rate represents the proportion of positive class errors predicted by the model to negative class, reflecting the error rate of the classifier in the positive class. The false negative rate formula is shown in equation (12).

$$FNR = \frac{FN}{FN + TP} \tag{12}$$

4.3 Evaluation Result

Using the 10-fold cross-validation method, three feature selection techniques were applied to two neural network models, and the experimental results were shown in Fig. 1 to Fig. 4.

Fig. 1 shows the experimental results of accuracy rate.

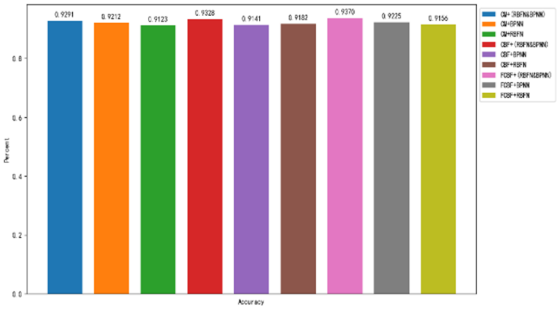

Fig. 1 Result of accuracy rate

According the analysis in Fig.1, FCBF has the best feature selection effect in the two neural network models applied to fusion (RBFN+BPNN) with an accuracy of 0.9370.

Fig. 2 shows the experimental results of recall rate.

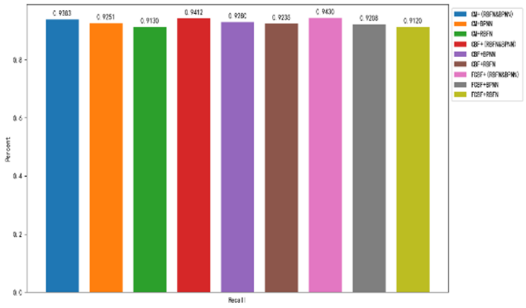

Fig. 2 Result of recall rate

As can be seen from the analysis in Fig.2, FCBF has the best feature selection effect in the two neural network models applied to fusion (RBFN+BPNN), with a recall rate of 0.9430.

Fig. 3 shows the experimental results of the false positive rate.

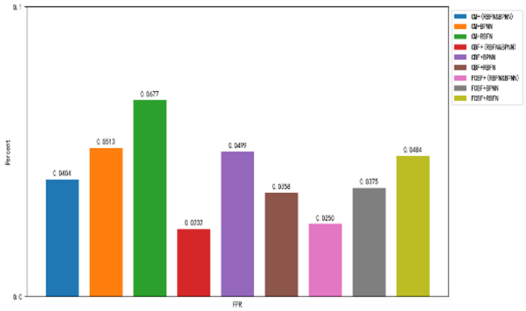

Fig. 3 Result of FPR

According to the analysis in Fig.3, FCBF has the lowest feature false positive rate in the two Neural Network models applied to fusion (RBFN+BPNN) with the FPR of 0.0250.

Fig. 4 shows the experimental results of false negative rate.

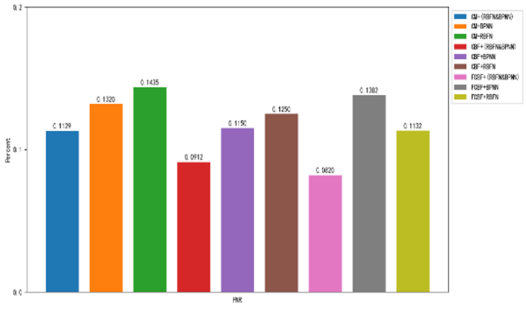

Fig. 4 Result of FNR

According to the analysis in Fig. 4, FCBF has the lowest feature omission rate in the two neural network models applied to fusion (RBFN+BPNN) with the FNR of 0.0820.

The comparative analysis of the experimental results of four evaluation indexes shows that FCBF has the best performance in the data feature selection of the two neural network models applied to fusion (RBFN+BPNN).

5. Conclusions and future work

Machine Learning techniques such as Neural Networks are well suited for Network Intrusion Detection, but the performance of classifiers can vary depending on the data set. Redundancy of data features is a major cause of variation in detection performance. In order to improve the detection performance, three feature selection techniques are discussed in this paper, and based on these feature selection techniques, a fusion method with RBFN and BPNN is proposed, and the actual network intrusion detection effect of various feature selection methods on KDD99 data set is evaluated through experiments. The experimental results show that when the feature selection method FCBF is combined with RBFN and BPNN integration methods, better detection results are produced.

There are two aspects of future research work. First, Network Intrusion Detection needs to process a large amount of data in practical applications, but this study is conducted on a relatively medium-sized network intrusion data set. In order to maintain the accuracy and efficiency of the classifier, it is necessary to study the scalability of this method when processing a larger data set, and explore how to effectively manage a large number of network intrusion data. Second, more data sets will be used to verify the effectiveness of feature selection methods, and more in-depth analysis and optimization of feature selection methods will be carried out to determine which features are most likely to be selected in network intrusion detection.

Acknowledgements

This research is partially supported by 2022 China University Research Innovation Fund - New Generation Information Technology Innovation Project(Grant No. 2022IT148), 2023 Chongqing Educational Science "14th Five-Year Plan" Special Key Topics for Education Reform(Grant No. K23ZG3050063).

References

[1] Kanagavalli P , Karthikeyani V .INVESTIGATION OF INTRUSION DETECTION SYSTEM USING RANDOM FOREST, CART AND PROPOSED SECURE RANDOM FOREST ALGORITHMS (SRFA)[J].journal of theoretical and applied information technology, 2023, 101(3):1196-1204.

[2] Ahmad I , Haq Q E U , Imran M ,et al.An Efficient Network Intrusion Detection and Classification System[J].Mathematics, 2022, 10.DOI:10.3390/math10030530.

[3] Wang Y, Meng W, Li W, Liu Z, Liu Y, Xue H. Adaptive machine learning-based alarm reduction via edge computing for distributed intrusion detection systems. Concurrency Comput: Pract Exper. 2019;31(19):1-12

[4] Kumar N , Kumar U .Artificial intelligence for classification and regression tree based feature selection method for network intrusion detection system in various telecommunication technologies[J].Computational intelligence, 2024, 40(1):e12500.1-e12500.23.DOI:10.1111/coin.12500.

[5] Krishnan V G , Saradhi M V V ,Lakshmi S.V.Kaviarasan S.Geetha A.NETWORK INTRUSION DETECTION BASED ON ONEDIMENSIONAL CNN WITH CHIMP OPTIMIZATION ALGORITHM[J].journal of theoretical and applied information technology, 2023, 101(10):3739-3748.

[6] Zolbayar B E , Sheatsley R , Mcdaniel P ,et al.Generating Practical Adversarial Network Traffic Flows Using NIDSGAN[J]. 2022.DOI:10.48550/arXiv.2203.06694.