Effectiveness Evaluation of Random Forest, Naive Bayes, and Support Vector Machine Models for KDDCUP99 Anomaly Detection Based on K-means Clustering

Majun Zhang

College of Electronic and Information Engineering, Tongji University, 201804 Shanghai, China

Abstract. Security in the World Wide Web has recently seen an enormous upgrade in almost every aspect. Identifying malicious activities in a network such as network attacks and malicious users plays a significant role in these upgraded security directions. This research utilizes the KDDCUP99 dataset to incorporate K-means clustering with three classifiers: Random Forest (RF), Naive Bayes (NB), and Support Vector Machine (SVM) with the goal to boost the accuracy of predicting network intrusions. In this paper, K-means clustering technique is applied as a preprocessing step to enhance the overall quality of network intrusion detection and maximize the accuracy of the network security measures. The goal is to identify anomalies with high accuracy. Experimental results indicate that the optimal combination is K-means + RF, which outperformed the others in precision, recall, and F1-score. Although K-means + NB demonstrated superior recall for certain smaller anomalies, it underperformed compared to the RF model. The paper concludes by highlighting the value of ensemble approaches, in particular Random Forest, for tackling anomaly detection and network security issues, particularly in light of the expanding significance of social networks and the internet.

1 Introduction

Protecting ever-expanding digital networks is a priority, given the cybersecurity risk landscape. This can be done through technical means — for example, anomaly detection dealing with unusual patterns in systems which may suggest a cyberattack or that are out of the ordinary.

In existing research, some common machine learning methods have been used for anomaly detection analysis on the KDDCUP99 dataset. O. E. Taylor combined principal component analysis (PCA) with RF [1], and references [2,3,4] utilized classic machine learning methods such as SVM, NB, and k-nearest neighbor. Many studies have also applied machine learning methods to different datasets such as UNSW-NB15, NSL-KDD or CICIDS-2017 for anomaly detection [5,6,7].And anomaly detection can be applied to not

Corresponding author: 2251335@tongji.edu.cn

only a variety of supervised machine learning models, but also some unsupervised techniques [8]. However, these models are not good at handling high-dimensional complex data and hence to detect anomalies become difficult.

In order to tackle this, a favourable way could be combining unsupervised learning with supervised methods [9,10]. The unsupervised K-means clustering technique can be used as a prior step to group data points by similarity, where this pre-computed similarity of clusters might help other models like RFs, NB and SVM in defining better decision boundaries. This study investigated whether this combination performs any better anomaly detection and in what scenarios.

2 Methodology

2.1 K-means clustering

K-means-clustering is an unsupervised technique to generate K number of clusters based on similarity into single cluster. It then assigns each sample to the nearest cluster and iteratively updates the centroid of it until convergence. The basic steps are:

1. Choose K initial cluster centers.
2. For every data sample, assign it to the center that is nearest.
3. The recalculation of cluster's centers by the mean value of samples.

Subsequent steps continue to iterate this process until the configuration of clustering no longer changes, which makes K-means much faster and feasible for large datasets.

2.2 SVM

SVM is a widely used classification method that separates data into two categories using an optimal hyperplane. The objective is to enhance the distance between the different categories, thereby bolstering the model's ability to generalize and minimize overfitting.

2.3 NB

Naive Bayes classifiers operate under the premise of Bayes' theorem and presume that the features are conditionally independent of each other, an assumption that streamlines the computational process. Despite this assumption, NB classifiers can often achieve commendable performance, particularly when applied to simpler datasets. This demonstrates their efficiency and adaptability in specific situations.

2.4 RF

RF is an ensemble technique that uses bootstrapped samples to construct multiple decision trees. In classification tasks, the prediction is based on the majority vote among the trees, while in regression tasks, the final prediction is obtained by averaging the individual forecasts from each tree. Each tree in the ensemble is trained on a random subset of features, contributing to the overall diversity and robustness of the model.

3 Experiment

This study uses the KDDCUP99 dataset, which originated from DARPA's 1998 intrusion detection project. The dataset contains records of TCP dump data and includes four types of

network anomalies: Probing, U2R, R2L, and DOS, with 41 features used for detection. Table 1 shows the labels for different type of abnormalities of the KDDCup99 intrusion detection datasets.

Table 1. KDDCup99 labels for Intrusion Detection Experimental Data

Label	Meaning	Specific Anomaly Types
Probing	Surveillance and other detection activities	ipsweep,nmap,portsweep,satan
U2R	Unauthorized entry by ordinary users into local superuser capabilities	Buffer_overflow,loadmodule,perl,rootkit
R2L	Unauthorized entry from remote devices	ftp_write,guess_passwd,imap,multihop,phf,spy,warezclient,warezmaster
DOS	denial-of-service attack	back,land,neptune,pod,smurf,teardrop
Normal	Normal activity	normal

The KDDCUP99 includes five million records in total, and this experiment uses a 10% training and testing subset provided by it. Table 2 shows the number of each label corresponding to the full dataset and the 10% data subset.

Table 2. Number of each label corresponding to the KDDCUP99 dataset

Class	Whole KDD	10% KDD
Probing	41102	4107
U2R	52	52
R2L	1126	1126
DOS	3883370	391458
Normal	972780	97278

3.1 Pre-Processing

Since the classification and regression model cannot handle non-numeric input variables and the 41 features provided by the KDDCUP99 dataset contain a total of four character features -- protocol_type, service, flag, and label, this paper first transforms them into numeric features. For example, if there is a normal value and four abnormal values in the label feature, the experiment will convert the normal character feature to 0, and the DOS character feature to 1, and the Probe to 2, etc. At the same time, there are some features with large value variations, such as src_bytes and dst_bytes in the range of about [0, 1379963888].And the experiment is to standardize and normalize these features in order to enhance the model's training speed and performance.

3.2 Kmeans processing

The experiment employs 41 factors other than labels as input features for Kmeans modeling, and it determines the number of clusters to be five depending on the types of labels in the dataset, then clusters using K-means and augments the original dataset with the output clustering results as a new feature.

3.3 Experiment results

After that, the clustering result of K-means is combined with the original 41 features to

form a new feature, which is used as the x-value for the subsequent model training, and the label that has been transformed into a numerical value is used as the target value y. The models are trained using RF, SVM, and NB models. After completing the training, these models are evaluated for test set effectiveness respectively. Table 3 presents the comparative results, showing that K-means combined with RF achieved the highest accuracy (99.97%), followed by SVM and NB.

Table 3. The effectiveness performance of each model

	Kmeans + NB	Kmeans + SVM	Kmeans + RF
Accuracy	0.880066	0.998549	0.999662

Fig. 1, Fig. 2, Fig. 3 shows the evaluation of the prediction effectiveness of the three models.

In the test data, the RF algorithm in conjunction with K-means is the most effective, SVM is in the middle, and NB is the least effective. Consistent with expectations, the overall effect of Naive Bayes on the complex training set is difficult to compare with classifiers such as RF and SVM, but it outperforms both of them on the recall test of U2R, which has the smallest number of samples, which suggests that when applied to small sample data, the NB model has a decent classification result. Furthermore, it appears that the RF method outperforms the SVM approach in terms of classification effect on this dataset, with accuracy, recall rate, and F1-score all appearing to be superior. At the same time, it can be seen that the shortcomings of the three models are that it is difficult to accurately classify the small dataset type of U2R.

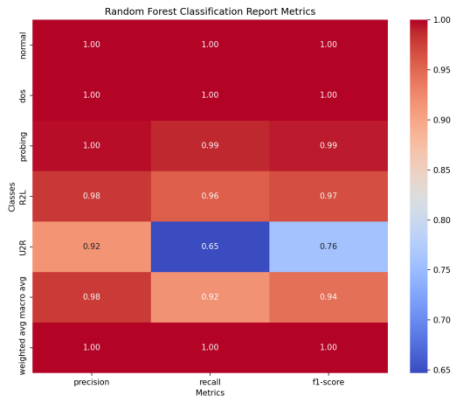

Fig. 1. RF Classification Report Metrics

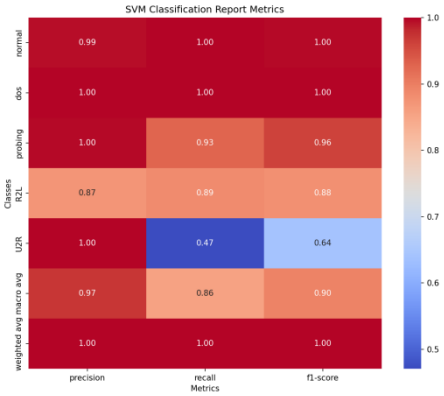

Fig. 2. SVM Classification Report Metrics

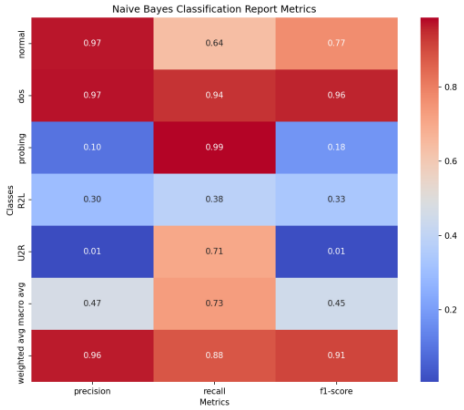

Fig. 3. NB Classification Report Metrics

4 Conclusion

Using the KDDCUP99 dataset, this paper evaluated the efficacy of K-means clustering in conjunction with RF, NB, and SVM for anomaly identification. The best accuracy and robustness were shown by K-means + RF, particularly for complicated, high-dimensional data. SVM performed well but was slightly less effective, while NB excelled in detecting smaller anomalies. These results suggest that combining clustering with ensemble learning techniques, such as RF, can significantly enhance anomaly detection. Future work could explore more advanced clustering and classification combinations, automated parameter optimization, and broader application to other cybersecurity datasets to verify the generalizability and scalability of the approach.

References

1. O. E. Taylor, P. S. Ezekiel, A smart system for detecting behavioural botnet attacks using random forest classifier with principal component analysis. European Journal of Artificial Intelligence and Machine Learning. **1**, 11 – 16 (2017)
2. I. Obeidat,N.Hamadneh,M.Alkasassbeh,M.Almseidin and M. AlZubi, Intensive Pre-Processing of KDD Cup 99 for Network Intrusion Classification Using Machine Learning Techniques. International Association of Online Engineering. **13**,70 (2019)
3. T. Mehmood and H. B. Md Rais, Machine learning algorithms in context of intrusion detection, 2016 3rd International Conference on Computer and Information Sciences (ICCOINS), (2016), 369-373
4. K. Limthong and T. Tawsook, Network traffic anomaly detection using machine learning approaches, 2012 IEEE Network Operations and Management Symposium, (2012), 542-545
5. M. Nawir, A. Amir, O.B. Lynn, N. Yaakob, and R.B. Ahmad, Performances of Machine Learning Algorithms for Binary Classification of Network Anomaly Detection System. J. Phys.: Conf. Ser. **1018** 012015 (2017)
6. B.S.Bhati,C.S.Rai,Analysis of Support Vector Machine-based Intrusion Detection Techniques. Arab J Sci Eng. **45**, 2371-2383 (2020)
7. N. Elmrabit, F. Zhou, F. Li and H. Zhou, Evaluation of Machine Learning Algorithms for Anomaly Detection, 2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), (2020), 1-8
8. A. M. Chandrashekhar, K. Raghuveer, Performance evaluation of data clustering techniques using KDD Cup-99 Intrusion detection data set. International Journal of Information and Network Security (IJINS). **1**, 294-305 (2012)
9. Y. Y. Aung and M. M. Min, An analysis of random forest algorithm based network intrusion detection system,18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), (2017), 127-132
10. Y. Y. Aung and M. M. Min, An analysis of K-means algorithm based network intrusion detection system, Advances in Science. Technology and Engineering Systems Journal. **3**, 496-501 (2018)