JURNAL RISET INFORMATIKA **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 4, No. 1 December 2021 DOI: https://doi.org/10.34288/jri.v4i1.147
Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

COMPARATIVE ANALYSIS OF THE K-NEAREST NEIGHBOR ALGORITHM ON VARIOUS INTRUSION DETECTION DATASETS

Andri Agung Riyadi [1)], Fachri Amsury [2)], Irwansyah Saputra [3)], Tiska Pattiasina [4)], Jupriyanto [5)]

Sains Data[1], Sistem Informasi[2,3,5]
Universitas Nusa Mandiri
andriagu1603@nusamandiri.ac.id[1], fachri.fcy@nusamandiri.ac.id[2], irwansyah.iys@nusamandiri.ac.id[3], jupriyanto.kahar@gmail.com[5]

Teknologi Informasi
Universitas Bina Sarana Informatika
elleoratiska07@gmail.com[4]

Abstrak

Keamanan di dalam jaringan komputer dapat menjadi rentan, hal ini disebabkan kita memiliki kelemahan dalam membuat kebijakan keamanan, konfigurasi sistem komputer yang lemah atau bug pada perangkat lunak. Intrusion detection adalah mekanisme mengamankan jaringan komputer dengan cara mendeteksi, mencegah, dan menghalangi usaha ilegal untuk mengakses informasi yang bersifat rahasia. Mekanisme IDS dirancang untuk dapat melindungi sistem dan mengurangi dampak kerusakan yang ada dari setiap serangan di dalam jaringan komputer karena melanggar kebijakan keamanan komputer meliputi ketersediaan, kerahasiaan, dan integritas. Teknik data mining telah digunakan untuk memperoleh pengetahuan yang berguna dari penggunaan dataset-dataset IDS. Beberapa dataset IDS yang umum digunakan adalah NSL-KDD, 10% KDD, Full KDD, Corrected KDD99, UNSW-NB15, ADFA Windows, Caida, dan UNM telah digunakan untuk mendapatkan tingkat akurasi menggunakan algoritma k-Nearest Neighbors (k -NN). Dataset IDS terbaru yang disediakan oleh Canadian Institute of Cybersecurity yang berisi sebagian besar skenario serangan terbaru bernama dataset CICIDS2017. Eksperimen pendahuluan menunjukkan bahwa pendekatan menggunakan metode k-NN pada dataset CICIDS2017 berhasil menghasilkan nilai rata-rata akurasi deteksi intrusi tertinggi dibandingkan dataset IDS lainnya.

Kata kunci: Intrusion Detection System, k-Nearest Neighbors, Machine Learning, Network Security

Abstract

Because we have flaws in developing security rules, inadequate computer system settings, or software defects, security in computer networks can be vulnerable. Intrusion detection is a computer network security method that detects, prevents, and blocks unauthorized access to confidential information. The IDS method is intended to defend the system and minimize the harm caused by any attack on a computer network that violates computer security policies such as availability, confidentiality, and integrity. Data mining techniques were utilized to extract relevant information from IDS databases. The following are some of the most widely utilized IDS datasets NSL-KDD, 10% KDD, Full KDD, Corrected KDD99, UNSW-NB15, ADFA Windows, Caida, dan UNM have been used to get the accuracy rate using the k-Nearest Neighbors algorithm (k-NN). The latest IDS dataset provided by the Canadian Institute of Cybersecurity contains most of the latest attack scenarios named the CICIDS2017 dataset. Preliminary experiment shows that the approach using the k-NN method on the CICIDS2017 dataset successfully produces the highest average value of intrusion detection accuracy than other IDS datasets.

Keywords: Intrusion Detection System, k-Nearest Neighbors, Machine Learning, Network Security

INTRODUCTION

The number of internet users around the world has exploded in the previous two decades. Hundreds of thousands of institutions and millions of people communicate with each other every day over the internet. As a result of these advancements, the number of attacks on internet networks continues to rise on a daily basis. Data integrity and privacy become a significant concern. The three

principles of network security are confidentiality, integrity, and availability, and network security attempts to defend the network from assaults on these three principles. An attempt to violate these three key characteristics is referred to as a network attack (Bace & Mell, 2001).

There is a lot of software that protects data and networks from incoming threats, such as firewalls, antivirus, data encryption, and user authentication, but it can't protect against all attacks. A lot of studies have been done on this subject to tackle this problem. Intrusion Detection Systems (IDS) was created to track and filter network activity by detecting threats and alerting network administrators (Chung & Wahid, 2012). The misuse detection method and the anomaly detection method are the two basic approaches for IDS. Ineffective against all forms of threats, yet each has its own set of strengths and weaknesses (Lin, Ke, & Tsai, 2015). Misuse detection is a methodical strategy to detect an assault on a computer network by comparing actions or looking for patterns that have previously been designated as attack symptoms. The abuse detection method is effective for detecting known assaults, but it is unable to detect fresh attacks (Zhang, Li, Gao, Wang, & Luo, 2015). Anomaly detection is useful at identifying novel assaults, with the exception that it is not very effective at known detection rates, resulting in a high FPR (Kim, Lee, & Kim, 2014).

Data mining techniques have been used to obtain useful knowledge from the use of IDS datasets. Some IDS datasets that are commonly used are NSL-KDD, 10% KDD, Full KDD, Corrected KDD99, UNSW-NB15, ADFA Windows, Caida, dan UNM have been used to get the accuracy using the k-NN algorithm approach (Hamid, et al., 2018). CICIDS2017, one of the latest IDS datasets from the Canadian Cybersecurity Institute (CIC) at New Brunswick University (UNB), was analyzed for research purposes (Sharafaldin, Habibi Lashkari, & Ghorbani, 2018). The CICIDS2017 dataset is created using a modern framework that takes into account your organization's policies and conditions and uses coefficients that can be individually determined for each criterion (Gharib, Sharafaldin, Lashkari, & Ghorbani, 2016).

The solution to overcome the challenges of fraud detection and anomaly detection technologies and maximize the capabilities of the two technologies is to use a hybrid approach(Depren, Topallar, Anarim, & Ciliz, 2005). For use with IDS, three hybrid methods are recommended: fraud detection and subsequent anomaly detection, anomaly detection and subsequent fraud detection, or fraud detection and anomaly detection at the same time. The IDS hybrid method uses a combination of many results from independent training of fraud detection and anomaly detection. For example, in the hybrid method, if at least one of the two methods classifies network traffic as an attack, then network traffic is considered an attack. In this case, the detection rate is high, but the IDS's FPR is still high. Conversely, if the hybrid method considers network traffic as an attack only if both methods are classified as attacks, the FPR will be low, but many attacks in the network traffic will be ignored (Kim et al., 2014). False Positive Rate (FPR) is when the IDS system detects benign or normal activity on the computer network and classifies it as a dangerous attack.

This research uses the K-NEAREST NEIGHBOR algorithm approach to measure the attack detection accuracy of the CICIDS2017 dataset. The algorithm method is not used in the CICIDS2017 dataset.

RESEARCH METHODS

In conducting research, will use Knowledge Discovery in Databases (KDD) method consisting of five stages, namely Data Selection, Preprocessing, Transformation, Data Mining, Interpretation, or Evaluation (Fayyad, 1997). The CICIDS2017 dataset will be used as the latest standard dataset for research and evaluation studies in the field of IDS, performing analysis to further identify the data, creating the initial findings, and then evaluating the quality of the data. The CICIDS2017 dataset consists of 3.1 million records with 85 attributes, including one attribute used as a label. Dataset attributes have seven attack categories and one normal category. The preprocessing process includes removing duplicate data, checking for inconsistent data, removing low-value or completely useless features, converting labels of all attack types to ATTACK labels, and fixing data errors. Feature selection is used to determine which features are important and to discard low-quality and uncorrelated features. Given the number of records in the CICIDS2017 dataset, you should perform data sampling for efficiency reasons. In this study, we obtained a 1% sample from the CICIDS2017 dataset. The fitted model is used to compare the result of the precision value using the k-NN algorithm approach with the value of $k = 5, 6, 7, 8, 9$. Results obtained in the form of accuracy, precision, and recall values are produced by comparison with other IDS datasets.

Literature Study

A. Intrusion Detection Systems and CICIDS2017 Dataset

Intrusion Detection Systems (IDS) are a very important part of protecting information systems in computer networks. The research report written by Anderson (1980) whose purpose was to enhance the audit capabilities of computer security and customer surveillance capabilities of the system, served as the initial concept of IDS (Anderson, 1980). There are three commonly used approaches to IDS systems: misuse detection, anomaly detection, and hybrid detection (McHugh, Christie, & Allen, 2000). Hybrid detection uses the IDS approach. This method combines the use of misuse detection and anomaly detection to improve the ability of both attack detection methods. There are three ways to implement hybrid methods on IDS, namely the use of the misuse detection method followed by the anomaly detection method, the anomaly detection method followed by the method of misuses detection or integrating the method of misuse detection, and the anomaly detection method at the same time (Kim et al., 2014).

CICIDS2017 is a dataset made by the University of New Brunswick's (UNB) Canadian Institute for Cybersecurity (CIC). CICIDS2017 was created using a modern framework that takes into account organizational policies and conditions with coefficients that can be individually determined for each criterion. This dataset consists of approximately 3.1 million records with more than 80 attributes, where 1 attribute is used as a label. The attributes in the dataset have 7 attack categories and 1 benign or normal category. The 7 categories of attacks in this dataset are Heartbleed Attack, Botnet, DoS Attack, Brute Force Attack, DDoS Attack, Infiltration Attack, and Web Attack.

B. Methodology

Data mining is the application of special algorithms to extract patterns from data (Fayyad, 1997). Data mining is about solving problems by analyzing existing data in the database (Witten et al., 2005). There are numerous techniques and methods for carrying out various types of data mining tasks. This method is classified into three major data mining paradigms, which are as follows: Predictive Modeling, Discovery, and Deviation Detection. Data mining and Knowledge Discovery in Databases (KDD) are frequently used interchangeably to describe the process of uncovering hidden information in a large database (Agushinta & Irfan, 2008). Although they have different concepts, data mining and KDD are related. Data mining is one of the stages in the KDD process.

One of the ten best data mining techniques is the k-Nearest Neighbors (k-NN) classification algorithm method. The k-Nearest Neighbors (k-NN) method uses famous Ciceroprinciplepares cum paribus facility congregant (birds of a feather flock together or equals with equals easily associate) (Mucherino, Papajorgji, & Pardalos, 2009). The accuracy, precision, and recall of eight different IDS datasets were compared using the k-NN algorithm. In the NSL-KDD dataset, the k-NN algorithm outperforms other algorithms in terms of accuracy, precision, and recall (Hamid et al., 2018).

The k-NN method does not generate a classifier from the data in a training set, but rather uses the training set every time classification is required; thus, the k-NN method is often referred to as a lazy classifier. Classification employs the analogy-based k-NN algorithm method, which compares test records with training records that have similarities. The k-Nearest Neighbors (k-NN) algorithm is a method for classifying objects that are based on learning data that is close to the object. This technique is very simple and straightforward to use. Similar to clustering techniques, namely grouping new data based on its distance from some existing data or its nearest neighbor. The first step is to calculate the distance to a neighbor before searching for data. Then, to define the distance between two points, namely the point on the training data and the point on the testing data, the Euclidean formula is used with equation (1), as follows:

$$d(a,b) = (x+a)^n = \sum_{i=0}^{n}(Xi - Yi)^2 \quad \text{..... (1)}$$

Explanation:
x: data 1
y: data 2
i: feature n-
d (a,b): Euclidean distance
n: number of features

In the concept of data mining, a confusion matrix is a method that is commonly used to calculate accuracy. If the dataset only has two classes, one is considered positive and the other is considered negative.

Table 1. Confusion Matrix

Data Class	Positive	Negative
Positive	true positives (TP)	false negatives (FN)
Negative	false positive (FP)	true negative (TN)

Accuracy is defined as the degree of similarity between predicted and actual values. Precision is the degree of agreement between the

information requested by the user and the response provided by the system. Precision values are calculated by dividing the number of positive examples correctly classified by the number of positive examples labeled as positive by the system. The recall rate is the system's success rate in rediscovering information. The recall value is calculated by dividing the number of correctly classified positive samples by the number of positive examples in the data.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

Research Framework Development

The ideas for this research were proposed using references from previous evaluations regarding data collection in several IDS datasets, beginning with the use of the CICIDS2017 dataset as a standard dataset for researchers in the field of Intrusion Detection Systems (IDS) (Sharafaldin et al., 2018). Data will be preprocessed by removing features that have been written twice, discarding wrong notes, no value, or incomplete data, and so on (Alshammari & Nur Zincir-Heywood, 2007; Radford, Richardson, & Davis, 2018). The dataset's data is labeled with two labels: BENIGN and ATTACK. The benign traffic represents normal network traffic, while the rest is an attack. It is feared that data cleansing or feature selection, which eliminates features that are less valuable or completely useless, will render research results irrelevant. To prepare for the data mining process, data reduction and data splitting are performed. The k-Nearest Neighbors (k-NN) approach will be used as a data mining algorithm to improve intrusion detection accuracy with values of k = 5, 6, 7, 8, 9.

RESULTS AND DISCUSSION

The data mining algorithm method approach is used at the classification stage to determine the accuracy of attack detection in the CICIDS2017 dataset. The algorithm used is the k-NN algorithm with values of k = 5, 6, 7, 8, 9.

A. K-Nearest Neighbour

Figure 1 depicts the accuracy, precision, and recall values in the CICIDS2017 dataset using the k-NN algorithm with values of k = 5, 6, 7, 8, 9.

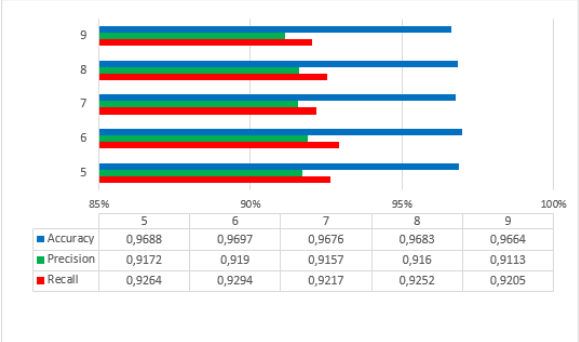

Figure 1. Classification Results Using the k-NN Algorithm

B. Comparison With Another Dataset

Table 2 compares the accuracy, precision, and recall values in several other IDS datasets using the k-NN algorithm with the value of k = 5, 6, 7, 8, 9.

Table 2. Comparison of the results of the k-NN algorithm on various IDS datasets

Dataset	Neighborhood														
	5			6			7			8			9		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Full KDD99	0.7342	0.722	0.734	0.70979	0.721	0.71	0.72028	0.736	0.72	0.65734	0.624	0.657	0.6958	0.633	0.696
Corrected KDD	0.6682	0.67	0.668	0.71495	0.707	0.715	0.48598	0.496	0.486	0.71962	0.722	0.72	0.71028	0.701	0.71
NSLKDD	0.7853	0.677	0.785	0.92	0.92	0.92	0.97592	0.959	0.976	0.9875	0.977	0.988	0.77193	0.77	0.772
10% KDD	0.8421	0.874	0.842	0.57142	0.571	0.571	0.64285	0.629	0.643	0.71428	0.706	0.714	0.5	0.521	0.5
UNSW	0.4285	0.351	0.429	0.57142	0.571	0.571	0.66083	0.655	0.661	0.8421	0.874	0.842	0.82456	0.83	0.825
Caida	0.6428	0.413	0.643	0.42857	0.762	0.351	0.5	0.521	0.5	0.71428	0.706	0.714	0.64285	0.413	0.643
ADFA Windows	0.7142	0.714	0.714	0.64285	0.413	0.643	0.82456	0.83	0.825	0.91228	0.92	0.912	0.85308	0.858	0.853
UNM Dataset	0.6382	0.626	0.638	0.79906	0.794	0.799	0.66822	0.67	0.668	0.72429	0.712	0.724	0.57943	0.53	0.579
CICIS2017	0.9688	0.9172	0.926	0.9697	0.919	0.929	0.9676	0.9157	0.922	0.9683	0.916	0.925	0.9664	0.9113	0.921

The k-NN algorithm with the value of k= 5, 6, 7, 8, 9 is used in Figure 2 to calculate the average value of accuracy, precision, and recall.

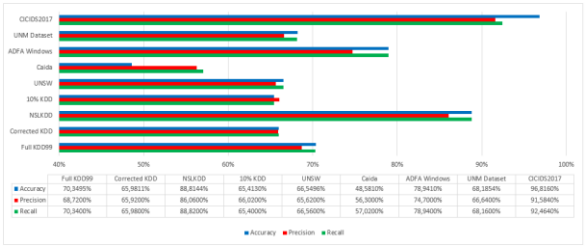

Figure 2. Comparison of the Average Values of Accuracy, Precision, and Recall of the k-NN Algorithm on Various IDS Datasets

We represented CICIDS2017 for comparison with several other existing IDS datasets; as shown in Figure 1, the highest accuracy value of the CICIDS2017 dataset was obtained using the k-NN algorithm with the value of k = 6, which equals 96.97%. Table 2 shows the accuracy, precision, and recall values in some IDS datasets, with the NSLKDD dataset having the highest level of accuracy using the k-NN algorithm with k = 8.

CONCLUSIONS AND SUGGESTIONS

Conclusion

The goal of this study is to detect network anomalies using machine learning methods. Because of its up-to-dateness, wide attack diversity, and various network protocols, the CICIDS2017 dataset was used in this context. The average value of accuracy, precision, and recall uses the k-NN algorithm with the value of k= 5, 6, 7, 8, 9 on the CICIDS2017 dataset higher than other IDS datasets which are 96.8160%, 91.5840%, 92.4640% as seen on Figure 2.

Suggestion

Based on the conclusions obtained, several suggestions can later be done for future research that researchers can use a more varied algorithm and up-to-date IDS datasets.

REFERENCES

Agushinta, D. (2008, August). Perancangan Aplikasi Data Mining Untuk Memrediksi Permintaan Customer Pada Perusahaan Persewaan Mobil. In *Proceeding, Seminar Ilmiah Nasional Komputer dan Sistem Intelijen (KOMMIT 2008)*.

Alshammari, R., & Nur Zincir-Heywood, A. (2007). *A flow-based approach for SSH traffic detection*. In *2007 IEEE International Conference on Systems, Man and Cybernetics* (pp. 296–301). IEEE. doi:10.1109/ICSMC.2007.4414006

Anderson, J. P. (1980). Computer security threat monitoring and surveillance. *Technical Report, James P. Anderson Company*.

Chung, Y. Y., & Wahid, N. (2012). A hybrid network intrusion detection system using simplified swarm optimization (SSO). *Applied Soft Computing*, *12*(9), 3014–3022. doi:10.1016/j.asoc.2012.04.020

Data Mining: Practical Machine Learning Tools and Techniques. (2011). Elsevier. doi:10.1016/C2009-0-19715-5

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & DATA, M. (2005). Practical machine learning tools and techniques. In *DATA MINING* (Vol. 2, p. 4).

Depren, O., Topallar, M., Anarim, E., & Ciliz, M. K. (2005). An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks. *Expert Systems with Applications*, *29*(4), 713–722. doi:10.1016/j.eswa.2005.05.002

Fayyad, U. (1997). *Data mining and knowledge discovery in databases: implications for scientific databases*. In *Proceedings. Ninth International Conference on Scientific and Statistical Database Management (Cat. No.97TB100150)* (pp. 2–11). IEEE Comput. Soc. doi:10.1109/SSDM.1997.621141

Gharib, A., Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2016). *An Evaluation Framework for Intrusion Detection Dataset*. In *2016 International Conference on Information Science and Security (ICISS)* (pp. 1–6). IEEE. doi:10.1109/ICISSEC.2016.7885840

Kim, G., Lee, S., & Kim, S. (2014). A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. *Expert Systems with Applications*, *41*(4), 1690–1700. doi:10.1016/j.eswa.2013.08.066

Lin, W.-C., Ke, S.-W., & Tsai, C.-F. (2015). CANN: An intrusion detection system based on combining cluster centers and nearest neighbors. *Knowledge-Based Systems*, *78*, 13–21. doi:10.1016/j.knosys.2015.01.009

McHugh, J., Christie, A., & Allen, J. (2000). Defending Yourself: The Role of Intrusion Detection Systems. *IEEE Software*, *17*(5), 42–51. doi:10.1109/52.877859

Bace, R., & Mell, P. (2001). *NIST special publication on intrusion detection systems*. Booz-allen and Hamilton Inc MCLEAN VA.

Mucherino, A., Papajorgji, P. J., & Pardalos, P. M. (2009). k-Nearest Neighbor Classification (pp. 83–106). doi:10.1007/978-0-387-88615-2_4

Radford, B. J., Richardson, B. D., & Davis, S. E. (2018). Sequence aggregation rules for anomaly detection in computer network traffic. *arXiv preprint arXiv:1805.03735*.

Sharafaldin, I., Habibi Lashkari, A., & Ghorbani, A. A. (2018). *Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization*. In *Proceedings of the 4th International Conference on Information Systems Security and Privacy* (pp. 108–116). SCITEPRESS - Science and Technology Publications. doi:10.5220/0006639801080116

Hamid, Y., Balasaraswathi, V. R., Journaux, L., & Sugumaran, M. (2018). Benchmark Datasets for Network Intrusion Detection: A Review. *Int. J. Netw. Secur.*, *20*(4), 645-654.

Zhang, J., Li, H., Gao, Q., Wang, H., & Luo, Y. (2015). Detecting anomalies from big network traffic data using an adaptive detection approach. *Information Sciences*, *318*, 91–110. doi:10.1016/j.ins.2014.07.044