Feature Dependent Naïve Bayes For Network Intrusion Detection System

Panny Agustia Rahayuningsih\textsuperscript{1,}\textsuperscript{*}, Reza Maulana\textsuperscript{2}, Windi Irmayani\textsuperscript{3}, Dedi Saputra\textsuperscript{4}, Deasy Purwaningtias\textsuperscript{5}

\textsuperscript{1,2,3} Sistem Informasi Akuntansi, Universitas Bina Sarana Informatika, Indonesia
\textsuperscript{4,5} Sistem Informasi, Universitas Bina Sarana Informatika, Indonesia

E-mail: panny.par@bsi.ac.id

Abstract. The intrusion detection system is an important component that performs the analysis for. the problem arising from the IDS is a collection of data sets in a computer network. to increase the high level and low false positive level of approach with the learning machine in applied. The data mining algorithm used is Naïve bayes one of the most widely used algorithms in space due to its simplicity, efficiency and effectiveness. NB has high accuracy and speed when applied into the database with large data. However, the NB algorithm assumes independent attributes (free) and is very sensitive to the selection of many features that interfere with the performance or accuracy of the NB to be low but in practice, the possibilities of the feature are interrelated. The Feature Dependent Naïve Bayes (FDNB) method is an effective method used to solve existing problems in NB by computing features as pairs and creating dependencies between each other as well as by applying learning models implemented to cross-validation, Feature Selection and data steps preprocessing that gives better accuracy results. After testing with two models of Naïve bayes and FDNB, the results obtained from the Naïve Bayes algorithm resulted in an accuracy of 84.42%, while for FDNB and oversampling (CFS + GS) the accuracy was 94.58%, FDNB and oversampling (CFS + BFS) the accuracy value of 94.69%, FDNB and SMOTE (CFS + GS) and FDNB and SMOTE (CFS + BFS) has an accuracy value of 93.27%. For the average per attack type DOS attack shows the highest result for its accuracy value of 97.86% and U2R attack produces the best accuracy when classifying U2R with 93.80% accuracy, U-F size of 96.26% U2R can be considered as a very result nice. Because U2R attack is considered very dangerous.

1. Introduction

As a dangerous activity in the integration, confidentiality or availability of computer resources. Intrusion is a type of attack that tries to bypass the security mechanism of a computer system [1]. Intrusion Detection System (IDS) is an important component in monitoring and analysis to detect, prevent internet attacks either host-based or network-based and possibly react to dangerous activities associated with computer network systems [4]. The Network Intrusion Detection System (NIDS) does packet logging, analyzes real time traffic from the IP network and tries to find out if there is an intruder trying to enter the computer network system [3]. An important problem that occurs in network-based intrusion detection systems is that a lot of data is collected and collected from network users, resulting in low detection accuracy, high false positivity [3]. Besides, by using the existing intrusion detection approach, IDS focuses on the selection of features [1]. To get detection with a high detection rate and a low false-positive rate.
So, to overcome these problems the intrusion detection approach is based on a Machine Learning (ML) based approach introduced. Machine Learning (ML) is an area of Artificial Intelligence that is concerned with the design and development of algorithms that enable computers to learn with the help of data. And the optimal features need to be done by contracting unnecessary features so that it can result in reduced processing time and higher detection accuracy [4]. The last few decades Decision Tree (DT), Naïve Bayesian (NB) Neural Network (NN), Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), Fuzzy Algorithms and Genetic Algorithms have been widely proposed by researchers for detecting intrusion NSL-KDD is a set of data that is recommended to solve some of the problems inherent in the KDD ’99 data set [5]. by providing significant improvements, this NSL-KDD data set was selected for testing and classification training. The classification determines the detection level of accuracy and the low false-positive rate of attacks detected on IDS. In the classification process, there are many algorithms used to determine which algorithm is considered best in the classification process. Several studies that have conducted research using this IDS dataset include, Improving the Performance of Multi-class Intrusion Detection Systems using Feature Reduction [7], Performance Comparison of Intrusion Detection Systems and Applications of Machine Learning to Snort System [8], and Research conducted by Frans Hendrik Botes, Louise Leenen and Retha De La Harpe in a paper entitled Ant Colony Induced Decision Trees for Intrusion Detection [9].

Naive Bayes (NB) is a machine learning algorithm that is often applied in classifications because efficiency and effectiveness also have high accuracy and speed values that are applied to databases with big data [6]. NB is widely used in network intrusion detection system (IDS) because it can improve the efficiency and accuracy of network intrusion detection and low false-positive rates, however, Nb has a weakness because it assumes independent attributes and is very sensitive to a large selection of features so that it interferes with performance or NB accuracy is low so it is necessary to research to improve its performance. But, in practice, the possibility of features are interrelated. This study proposes a new method by following the steps of data preprocessing in the Feature Dependent Naïve Bayes (FDNB) method where FDNB will be applied to calculate features as pairs and create dependencies between each other and by applying learning models that are implemented into cross-validation, Feature Selection and following the steps of data preprocessing (normalization, discretization, class distribution balancing) which gives better accuracy results [10].

This FDNB method is applied to IDS problems by using the NSL-KDD dataset to detect network attacks in four main attack categories: Probes (information gathering), DoS (service denial), U2R (users for root) and R2L (users remote to local) [2] to determine which detection level is highest and false positives are low. The proposed method is also compared with standard NB classifiers [10]). This research is conducted to find out which ML algorithm is good in classifying NSL-KDD data on intrusion detection on computer network security. The results of the classification of each algorithm will be compared and seen the level of accuracy, precision, f-means, probability of detection (pd), and probability of false alarm (pf).

2. Method

Propose the FDNB method by following the data preprocessing steps that are applied to the learning model [10]. Learning models are built based on training data and no testing data is included. The learning model in this study is shown in Figure 1.

The requirement to create a learning model infrastructure is to create a learning model from training data not from test data [13]. This requirement is very important for the sake of compatibility with new data [10]. The following are the data preprocessing stages which apply training data where the parameter results obtained in this process will be used for preprocessing test data. In the formula for preprocessing training data classification becomes the determining variable. Where is the estimated class, each sample in the test data carried out using the
classification formula will then produce a performance report. In the formula for preprocessing training data classification becomes the determining variable. Where is the estimated class, each sample in the test data carried out using the classification formula will then produce a performance report. Next, each step must be repeated in accordance with the cross validation method M x N. In this method, all datasets must be divided into sets of N, namely one subset as test data and the rest (N-1) is used as training data. In order to enable each subset to be used as test data this process must be repeated N times. Thus, errors that arise from sample selection can be minimized. On the other hand, to get good results and randomized data set ranks at each iteration, this process must be repeated M times to reduce the effect of sample orders. In this study 200x10 cross validation was used MxN ie M = 20 and N 10. The results obtained from 200 learning models were then evaluated and compared with other studies.

2.1. Model Experiments and Testing

NSL-KDD is a set of data that is recommended to solve some of the problems inherent in the KDD ’99 data set [5] by providing significant improvements, this NSL-KDD data set was selected for testing and classification training. Data sets have also been processed before using categorical category encoding to convert categories to numeric values so that nominal fields are represented in numerical categories, not in the text. Nominal fields represent certain classes or categories, for example, Transmission Control Protocol (TCP) and User Datagram Protocol (UDP). The data set used in this study is available at http://github.com/FransHBotes/NSLKDD-Dataset. The task of classification idurs binary: normal or malicious traffic [11].

In the dataset, each record is 41 attributes, there are types of attacks and normal that contains information about NSL-KDD and types of attacks from all 41 attributes that exist in the NSL-KDD dataset. The 42rd attribute contains data about various 5 vector network connection classes and is categorized as one normal class and four attack classes.
2.2. Evaluation and Validation of Results
The preprocessing data stage is applied to the training data, while the parameters obtained from the training data will be processed for use in the preprocessing data testing data. In the learning model done using java language. Weka tool as one of the steps to implement is implemented from Cross-Validation, Feature Selection, and data preprocessing.

3. Results and Discussion
In Table 1 there are the results of the average value of all existing attacks and Figure 2 shows the results of the comparison of all models that have been tested. From this figure, it can be seen that the FDNB + OS (CFS + GS) and FDNB + OS (CFS + BFS) models are better models for improving accuracy, precision, f-measure, PD, and pf performance by obtaining average values which is not much different from obtaining the highest accuracy, precision, f-measure, low PD and pf values compared to the others.

| Model       | FDNB+OS (CFS+GS) | FDNB+OS (CFS+BFS) | FDNB+SMOTE (CFS+GS) | FDNB+SMOTE (CFS+BFS) |
|-------------|------------------|-------------------|---------------------|----------------------|
| acc         | 80.62            | 93.82             | 93.00               | 93.27                |
| precision   | 81.69            | 90.39             | 90.45               | 89.60                |
| f-mean      | 75.82            | 90.90             | 90.89               | 90.40                |
| pd          | 76.69            | 91.57             | 91.44               | 91.87                |
| pf          | 22.08            | 7.50              | 7.59                | 21.02                |

Figure 2.
Comparison diagram of the overall average

From table 2 above, the FDNB method classification for Dos and Probe attacks has the highest verification value, namely Dos attacks of 97.86 % and Investigations of 95.45 % with low pf levels of 1.25 % for Dos while Probe has a level pf which is quite high at 4.39 %. For F-size DOS 96.75 % and probes with 95.19 % values. An F-DNB classification was obtained which compiled well classifying U2R with an accuracy of 93.85 %, and an F-size value of 96.66 % could be considered an excellent result. Because U2R attacks are very dangerous. More clearly can be seen in the graph below. Following are the results of type per attack in Figure 3 below.
Table 2. Measurement Results of Avg. Attack Type

| Model  | DOS   | Probe  | R2L   | U2R   |
|--------|-------|--------|-------|-------|
| acc    | 97.86 | 95.45  | 89.30 | 93.85 |
| precision | 97.43 | 95.17  | 82.96 | 96.09 |
| f-mean | 96.75 | 95.19  | 88.71 | 96.66 |
| pd     | 96.08 | 95.26  | 95.52 | 97.29 |
| pf     | 1.25  | 4.39   | 15.57 | 43.44 |

Figure 3. Average Comparison Diagram Per Attack Type

4. Conclusion
In the measurement of accuracy, f-measure, precision, pd, and pf NB model shows a decrease compared to other models. The FDNB + OS (CFS + BFS) model shows an increase in accuracy of 94.69%, the FDNB + OS (CFS + GS) model does not differ much by 94.58%. While the other models have relatively the same or down accuracy, but the measurement of accuracy does not consider class imbalances. In measurements that consider class imbalances such as f-measure, pd / recall show a clear increase in the FDNB + OS (CFS + GS) model, FDNB + OS (CFS + BFS), FDNB + SMOTE (CFS + GS) and FDNB + SMOTE (CFS + BFS). Thus, the research problem, problem formulation, and research objectives are answered in this study that the FDNB method is better than the NB method in the NSL-KDD dataset. Then, the Test21 Dataset is used for test purposes because of unknown attack features. Unlike other test datasets, Test21 specifically excludes attacks that are easily detected because it can produce a performance of up to 65% [5]. For the average per attack type, Dos attack showed the highest results for its accuracy value of 97.86%, followed by the second for the probe of 95.45%, R2L with an accuracy of 89.30%, while the U2R attack produced the best accuracy when classifying U2R with an accuracy of 93.85%, with a very unbalanced dataset, an F-mean U2R size of 96.66% can be considered a very good result. Because U2R attacks are considered very dangerous [9]. DOS attack also shows a lower pf value of 1.25% compared to Probe, and R2L while U2R attacks show the highest pf value of 43.44%. Thus, the research problem, problem formulation, and research objectives are answered in this study that the FDNB method provides the best intrusion detection results.

Some research results might encourage future research. In order to improve this research, the following suggestions are proposed:

(1) It can focus on different preprocessing data so that it can analyze the performance of the
FDNB classification

(2) The results of research for the standard NB model do not do feature selection and data preprocessing so as to produce a lower value than the proposed FDNB model. Further research may be carried out by providing feature selection and data preprocessing for the NB method.

(3) Developing using other feature selection methods such as chi-square, information gain and gain ratio.

(4) Future studies can compare with different algorithm models such as Neural Network, K-NN, Decision Tree and other algorithms.

(5) Further research can use different datasets.

References

[1] Mukherjee S and Sharma N, 2012 Intrusion Detection using Naive Bayes Classifier with Feature Reduction Procedia Technol. 4 p. 119–128.
[2] Kumar K, 2016 Network Intrusion Detection with Feature Selection Techniques using Machine-Learning Algorithms 150, 12 p. 1–13.
[3] Harbi N and Rahman M Z, 2010 OMBINING N AIVE B AYES AND D ECISION T REE 2, 2 p. 12–25.
[4] Mohammad M N Sulaiman N and Muhsin O A, 2011 A novel Intrusion Detection System by using intelligent data mining in WEKA environment Procedia Comput. Sci. 3 p. 1237–1242.
[5] Tavallaee M Bagheri E Lu W and Ghorbani A A, 2009 A Detailed Analysis of the KDD CUP 99 Data Set Cisda p. 1-6.
[6] Wu J Pan S Zhu X Cai Z, Zhang P and Zhang C, 2015 Expert Systems with Applications Self-adaptive attribute weighting for Naive Bayes classification Expert Syst. Appl. 42, 3 p. 1487–1502
[7] Wahba E Elsalamouny E and Eltaweel G, 2015 Improving the Performance of Multi-class Intrusion Detection Systems using Feature Reduction Improving the Performance of Multi-class Intrusion Detection Systems using Feature Reduction July.
[8] Ali S Shah R and Issac B, 2017 machine learning to Snort system Futur. Gener. Comput. Syst.
[9] Hendrik F, 2017 Information Warfare December.
[10] Arar Ö F and Ayan K, 2017 A Feature Dependent Naive Bayes Approach and Its Application to the Software Defect Prediction Problem Appl. Soft Comput. J.
[11] Paper C and Council L L, 2017 Ant Colony Induced Decision Trees for Intrusion Detection June.
[12] Fayyad U M and Irani K B, 1993, Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning, Proceedings of the International Joint Conference on Uncertainty in AI. p. 1022–1027.
[13] Song Q Jia Z Shepperd M Ying S and Liu J, 2011 A General Software Defect-Proneness Prediction Framework 37, 3 p. 356–370.