An Intrusion Detection Model based on Hybrid Classification algorithm

Manfu Ma[1,2], Wei Deng[1,2], Hongtong Liu[1], Xinmiao Yun[1,2]

[1]College of Computer Science and Engineering, Northwest Normal University, Gansu, China
[2]Internet of Things Engineering Research Center, Gansu, China

Abstract. Due to using the single classification algorithm can not meet the performance requirements of intrusion detection, combined with the numerical value of KNN and the advantage of naive Bayes in the structure of data, an intrusion detection model KNN-NB based on KNN and Naive Bayes hybrid classification algorithm is proposed. The model first preprocesses the NSL-KDD intrusion detection data set. And then by exploiting the advantages of KNN algorithm in data values, the model calculates the distance between the samples according to the feature items and selects the K sample data with the smallest distance. Finally, by naive Bayes to get the final result. The experimental results on the NSL-KDD dataset show that the KNN-NB algorithm can meet the requirement of balanced performance than the traditional KNN and Naive Bayes algorithm in term of accuracy, sensitivity, false detection rate, specificity, and missed detection rate.

1 Introduction

At present, intrusion detection technology can be divided into two categories: anomaly detection and misuse detection[1]. Anomaly detection is a detection mode that detects an intrusion based on detecting an access that deviates from normal behavior. Misuse detection is a detection mode that discovers an intrusion based on a known attack type. Although accuracy is the basic requirement of IDS, its expandability and adaptability are also crucial in today's network computing environment[2]. The characteristics of data mining classification algorithms are adapted to the type judgment of unknown data by the type of known data[3]. It is in good agreement with the misuse detection principle in intrusion detection technology. There many data mining techniques applying to IDS. Among them, the literature [4-6] used several classification algorithms in data mining to analyze the KDD CUP 99 intrusion detection data set. Kulariya M et al used KNN algorithm to perform intrusion detection and the experimental results show that the KNN algorithm when doing intrusion detection has high accuracy and specificity and efficiency, but its false alarm rate and missed detection rate are high and the sensitivity is low[4]. This will cause many abnormal behaviors to be detected without harming the system or a lot of normal. Behavior is detected anomalous that wasting too much human and financial resources to deal with these normal behaviors. The literature [5-6] compares the performance of many classification algorithms in intrusion detection. The experimental results show that Naive Bayes has higher sensitivity and faster training time than other classification algorithms, but the accuracy and specificity of the performance is poor.

Therefore, based on the above issues, this paper fully understands and analyzes the principles of Naive Bayes and KNN classification algorithms, and then understanding and analyzing the intrusion detection data set, it proposes a algorithm based on KNN and Naive Bayes classification of the intrusion detection model. The model uses the advantage of KNN algorithm in data values, calculates the distance based on the feature items between the samples, and selects the K sample data with the smallest distance, thereby excluding the uncorrelated sample data and reducing the size of the sample data set; then using the advantage of Naive Bayes algorithm in data structure, the obtained K sample data are trained and the classifier is constructed. Finally, the posterior probability is obtained by the prior probability to obtain the final classification result. The experimental results on the NSL-KDD data set show that the KNN-NB algorithm can meet the performance equalization requirements in terms of accuracy, sensitivity, error rate, specificity and missing rate, etc., compared with traditional KNN and naive Bayesian algorithms.

The rest structure of this paper is as follows: The second part introduces the related contents of Naive Bayes and KNN algorithms. In the third part, the hybrid classification algorithm based on KNN and naive Bayesian KNN-NB is described in detail, and the algorithm is applied to the intrusion detection model. The fourth part introduces the used data sets, experimental tools and platforms, and conducts experimental analysis. Finally, we have a summary.

2 Related work

[a] Corresponding author: 806345788@qq.com

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

2.1 KNN algorithm

KNN algorithm, also known as K-nearest neighbor algorithm, is a classic data mining algorithm. The basic idea of the algorithm is that some phenomena in the real world can be reproduced under certain conditions, so long as the conditions in the past have been satisfied or similar, this phenomenon may occur[7].

The algorithm flow of KNN algorithm is very simple. Assume that the sample set size is n, m is the number of feature items, c is a category, and one of the test data X can be expressed as $X_i=\{x_1, x_2,..., x_{m-1}, x_m\}$, $i\in[1, n]$, a piece of sample data Y can be expressed as $Y_j=\{y_1, y_2,...,y_{m-1}, y_m, c_j\}$, $j\in[1, n]$. First, the data is preprocessed; then the distance between the measured data and each sample data in the sample set is calculated; then the K sample data closest to the data to be measured are selected from the sample set; finally, the category of the K sample data closest to the data to be measured is counted, and the category with the most categories is selected from among the categories, and this category is the category of the data to be measured. The Euclidean distance is generally used to calculate the distance between the measured data and the sample data. The Euclidean distance expression is shown in expression (1).

$$dis(X_i, Y_j) = \sqrt{\sum_{k=1}^{m} (x_m - y_m)^2} \quad (1)$$

From expression (1), it can be seen that the KNN algorithm is analyzed from the numerical aspects of the data to obtain the classification category.

2.2 Naive Bayes algorithm

Naive Bayesian algorithm is a Bayesian classification algorithm based on Bayesian theorem, the precondition of which is that each feature item is independent, and each feature item is equally important, which is also the simplicity of naive Bayesian. The core idea of naive Bayesian algorithm is to solve the posteriori probability by using the pre-test probability [8].

Naive Bayes algorithm is also a classic data mining algorithm, it has a solid foundation of classical mathematical theory, and has good stability in the classification effect[9]. The basic idea is: for a given test data $X_i = \{x_1, x_2, ..., x_{m-1}, x_m\}$, where x_m represents the feature item and m is the number of feature items, which is solved in the test. The probability of occurrence of each category c_j under the condition of occurrence of feature items in the data, and classify the test data into the category where $P(c_j|X)$ is the largest.

Definition 1 (Conditional Probability) The probability P(B|A) that event A occurs under the condition that event B occurs. The expression is:

$$P(A|B) = \frac{P(AB)}{P(B)} \quad (2)$$

Derivation of this expression gives the multiplication expression :

$$P(AB) = P(A|B)P(B) = P(B|A)P(A) \quad (3)$$

So expression(2) can be written as:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad (4)$$

Definition 2 (division of sample space) Let Ω be the sample space of test E, and $B_1, B_2,...,B_n$ be a set of events, if:

$$1、B_iB_j = \emptyset, i, j=1,2,...,n;$$

$$2、B_1 \cup B_2 \cup ... \cup B_n = \Omega$$

We call $B_1, B_2,...,B_n$ a division of the sample space Ω.

Definition 3 (Bayesian formula) Let Ω be the sample space of test E, A be the event of E, $B_1, B_2,...,B_n$ be a division of Ω, and P(A)>0, $P(B_i)$> 0 (i=1, 2,...,n), then:

$$P(B_i|A) = \frac{P(A|B_i)P(B)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}, i = 1,2,...,n \quad (5)$$

Call this the Bayesian expression.

Assuming that the test data X can be expressed as $X_i=\{x_1, x_2,..., x_{m-1}, x_m\}$, the sample set size is n, m is the number of feature items, and the category set C is represented as $C=\{c_1, c_2,...,c_{j-1}, c_j\}$(i, $j\in[1, n]$). In the Naive Bayes algorithm, assuming that the feature items are independent of each other, what is sought is the probability of occurrence of each category under the conditions in which the feature items appear in the data to be measured. The Bayesian expression can be expressed as:

$$P(c_j|X_i) = \frac{P(X_i|c_j)P(c_j)}{P(X_i)} = \frac{P(x_1|c_j)P(x_2|c_j)...P(x_m|c_j)P(c_j)}{P(X_i)} \quad (6)$$

As long as the category corresponding to the maximum value in $P(c_j|X_i)$ is found, the data to be measured is classified in this category. Since the value of $P(X_i)$ is the same for all classes of $P(c_j|X_i)$, we only need to compare $P(x_1|c_j)P(x_2|c_j)...P(x_mc_j)P(C_j)$ can be the size. When the number of samples is large, according to the central limit theorem, the frequency is equal to the probability. Therefore, $P(x_mc_j)$ and $P(c_j)$ can be obtained from the sample, and finally the type of the data to be measured can be determined.

3 Intrusion Detection Model Based on KNN-NB Algorithm

3.1 The basic idea and process of the algorithm

The KNN algorithm and the naive Bayes algorithm are introduced above. Although the KNN algorithm and naive Bayesian algorithm have their own advantages, the

shortcomings are obvious. For the KNN algorithm, it only from the data of the numerical aspects of category scoring, scoring is not normalized, and since the KNN algorithm is obtained from the K nearest neighbor sample data to obtain the final category, it will appear when the sample data imbalance(For example, if there is a very large number of samples in one category in the sample, and there are fewer samples in other categories), the final result will be biased to the category of the number of categories. And the KNN algorithm has poor noise resistance, and the classification result is biased due to excessive data noise. For naive Bayesian algorithm, it classifies the data from the structural aspect, calculates the posteriori probability by calculating the prior probability, but the precondition is that the characteristic items of the data are independent and the result of the classification can reach the ideal value [10]. But in the actual situation, this hypothesis is generally not tenable, and the Naive Bayesian algorithm has higher training complexity.

Based on the above problems, this paper proposes a hybrid classification algorithm KNN-NB based on KNN and Naive Bayes. Assume that the sample set size is n, m is the number of feature items, c is a category, and one of the test data X can be expressed as $X_i=\{x_1, x_2,..., x_{m-1}, x_m\}$, $i\in[1, n]$, a piece of sample data Y can be expressed as $Y_j=\{y_1, y_2,...,y_{m-1}, y_m, c_j\}$, $j\in[1, n]$.

The detailed steps of the KNN-NB algorithm are as follows:

Step 1. The sample set Y and the test set X data are preprocessed to make the data conform to the algorithm input format;

Step 2. Select the appropriate K value;

Step 3. Use expression (1)we need to calculate the distance between the measured data X_i and the sample set Y, and select the smallest K sample data;

Step 4. The selected K sample data as the Naive Bayesian algorithm sample set Y', and then using the formula (6) to calculate the data X_i and the sample set Y' class probability P (c_j| X_i);

Step 5. Finally,we classify the measured data X_i into the category with the highest probability P(c_j|X_i).

The KNN-NB algorithm flow is shown in Figure (1).

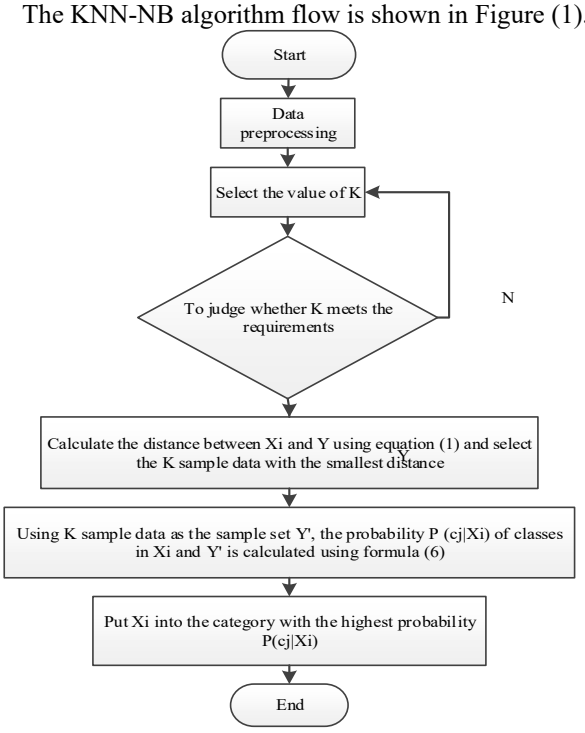

Figure 1. KNN-NB algorithm flow chart.

3.2 Intrusion Detection Model Based on KNN-NB Algorithm

In this paper, an intrusion detection model is established using KNN and Naive Bayes-based hybrid classification algorithm KNN-NB. The model architecture is shown in Figure 2. The model is mainly divided into three stages, each stage is described as follows:

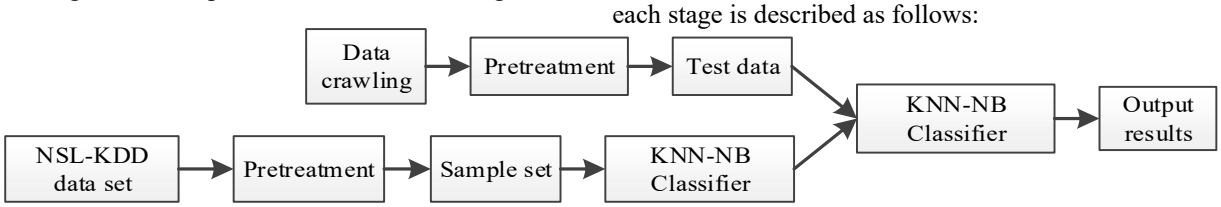

Figure 2. Intrusion detection model based on KNN-NB algorithm.

The first stage: data acquisition and preprocessing stages. The main work in this phase is the generation of measured data and sample sets. For the data to be measured, first use the packet grabber to grab the data packet and get rough data from it; then the data is processed into the format that meets the input of the next stage through preprocessing, and finally the data to be tested is generated. This phase takes up about 60% of the entire model time.

The second stage: KNN-NB training phase. The main task in this phase is to generate the KNN-NB classifier. The KNN-NB trainer first initializes a K value, obtains K sample data through the input feature item xm and sample set Y, and then uses the K sample data as a new sample set Y' to determine the prior probability, thereby generating KNN-NB classifier. This is a process of constantly repeating supervised learning. Through constant learning, the optimal K value and prior probability are determined to generate an optimal classifier.

The third stage: KNN-NB classification stage. This phase is the process of implementing intrusion detection. The measured data Xi is classified by the trained optimal classifier, and the final classification result is output, that is, whether or not the judgment belongs to an intrusion behavior.

4 Experiments and results analysis

4.1 Experimental data and preprocessing

4.1.1 Sample data

The experimental data in this paper is based on the NSL-KDD standard intrusion detection data set, which is generated after the optimization of the KDD-CUP 99 DataSet [11]. There is no duplicate data in the NSL-KDD test set, making the experimental accuracy more convincing. The proportion of normal and abnormal data in the NSL-KDD dataset and the amount of test and training data are reasonable, so it is more suitable for effective and accurate evaluation of different data mining algorithms. This paper uses 20% training dataset and test dataset in the NSL-KDD dataset for experiments. The dataset structure is shown in Table 1.

Table 1. Training set and test set structure.

Category	Training set	Ratio(%)	Test set	Ratio(%)
Normal	65536	53.16	9711	53.61
DOS	45927	37.26	5064	27.95
Probe	11656	9.46	1106	6.10
R2L	103	0.08	2199	12.14
U2R	52	0.04	37	0.20
Total	123274	100	18117	100

4.1.2 Sample Data Preprocessing

Before the experiment, the first task is to perform data preprocessing[12].After analyzing the NSL-KDD data set, it was found that the data set is 43-dimensional data, and the 43-dimensional data is the number of labels predicted by the community that created this data set when using different algorithms to perform experiments. This value is not useful in this experiment, so the dimension data is deleted. Therefore, the first 41 features and 42nd dimension items in the data set are used in this experiment. Moreover, the 2nd, 3rd, 4th, and 41th-dimensional data in the data set are discrete data, which does not conform to the input data format of the algorithm. Therefore, this paper will perform the data processing on this four-dimensional data, and the comparison table before and after processing is shown in Table 2. Show.

Table 2. Discrete data numerical comparison table.

Feature item	Before and after quantification
protocol_ty pe	icmp-1, tcp-2, udp-3
service	domain_u-1, ecr_i-2, eco_i-3, finger-4, ftp_data-5, ftp-6, http-7, hostname-8… OTH-1, REJ-2, RSTO-3, RSTOS0-4…
flag	normal-0, others-1
label	

4.1.3 Experimental Platform and Evaluation Standards

The experiment in this paper is done in a memory of 8GB, and the processor is an Intel Core(TM) i5-7200U CPU 2.50GHz. The system is implemented under the Windows 10 platform. The program uses a single thread implementation, and only uses one CPU core to run the calculation process.

This paper uses the general performance evaluation criteria of intrusion detection to compare the model proposed in this paper with the model using KNN algorithm and Naive Bayes algorithm. Performance evaluation is based on the following criteria[13].

Accuracy: $$AR = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

Specificity: $$SpeR = \frac{TN}{TN + FP} \quad (8)$$

Sensitivity: $$SenR = \frac{TP}{TP + FN} \quad (9)$$

Misuse rate: $$FR = \frac{FP}{FP + TN} \quad (10)$$

Missed rate: $$MR = \frac{FN}{TP + FP + FN + TN} \quad (11)$$

Among them, true positive (TP): Attack category is identified as the attack category. False Positive (FP): The normal category is identified as the attack category. True Negative (TN): The normal category is identified as the normal category. False Negative (FN): The attack category is identified as the normal category.

4.2 Analysis of Results

According to the above evaluation criteria, this article uses the literature[4] and the literature[5] with the KNN-NB algorithm proposed in this paper to conduct comparative experiments. The experiment adopts the same sample set and test set. The results of the experiment were averaged after three experiments.

The missed detection rate is an important indicator for evaluating an intrusion detection model, because when an attack behavior is missed as normal behavior, the terminal is likely to receive an attack, resulting in irreparable loss[14]. As can be seen from Figure 3, the KNN-NB algorithm in K<4250, the missed detection rate is lower than the KNN algorithm and the Naive Bayes algorithm, for example, when K=2500, the KNN-NB algorithm has a lower than expected rate than the KNN algorithm. The Naive Bayes algorithm decreased by 75.5% and 29.4% respectively.

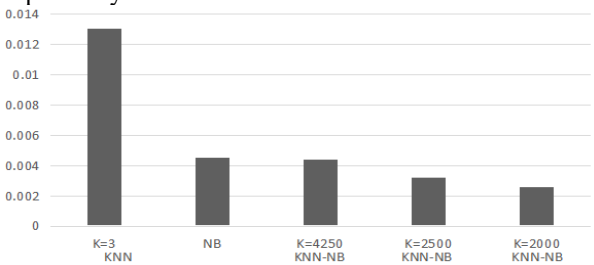

Figure 3. Comparison of missed detection rates for different algorithms.

The false detection rate is the ratio of how many normal categories are misclassified as the attack category with respect to the total amount of the normal category. The lower false detection rate, the higher credibility of the model, the higher false detection rate of the model will make the detection personnel mistaken for this behavior is an aggressive behavior, will take corresponding measures to deal with, resulting in useless manpower loss[15]. The experiment compares the false detection rate of KNN algorithm, Naive Bayes algorithm and KNN-NB

algorithm, as shown in Figure 4. As can be seen from the diagram KNN-NB algorithm when K=4250, the detection error rate is less than the Naive Bayes algorithm, and far less than the KNN algorithm, error detection rate by 23.5% and 65.3 respectively, so as to make the terminal will not mistakenly identified under attack because of the detection system.

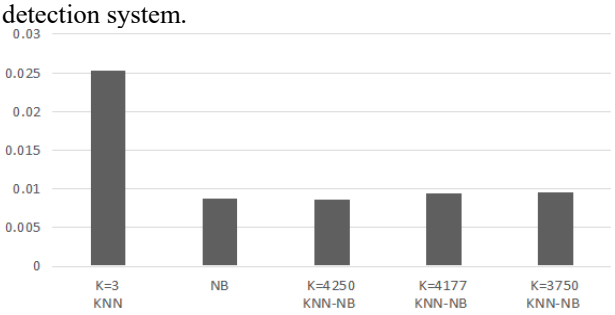

Figure 4. Comparison of false detection rates for different algorithms.

Sensitivity is a measure of the type of attacks that are truly classified, and is an important indicator of whether a model is good or bad. Only the type of attack is accurately detected and the model has value to use[16]. As can be seen from Figure 5, the KNN-NB algorithm has the best sensitivity at K=4250, and it was found in the experiment that when K is in the range [3750, 4250], the sensitivity is better than that of the KNN algorithm and the Naive Bayes algorithm.

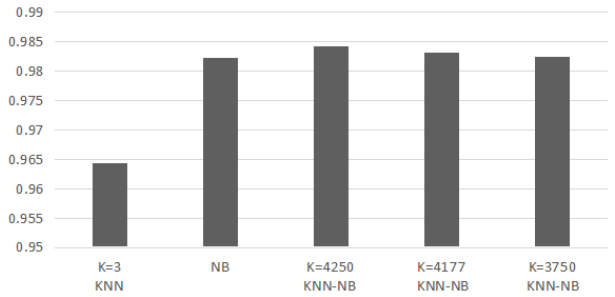

Figure 5. Comparison of sensitivity of different algorithms.

The accuracy and specificity of the experimental results are shown in Figure 6. It can be seen from Fig. 6 that the accuracy and specificity of the algorithm KNN-NB is lower than that of the KNN algorithm. However, within the acceptable range, and compared with Naive Bayes algorithm, the accuracy and specificity of KNN-NB algorithm is relatively high.

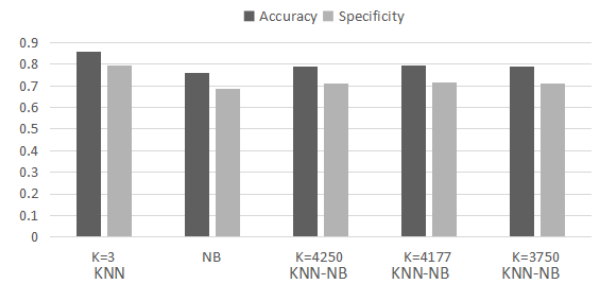

Figure 6. Comparison of accuracy and specificity of different algorithms.

5 Concluding remarks

This paper proposes an intrusion detection model based on KNN and Naive Bayes hybrid classification algorithm. This model balances the problem of performance imbalance with a single classifier by using the advantages of KNN algorithm in data values and the advantages of Naive Bayes algorithm in data structure. The experimental results show that the intrusion detection model based on KNN-NB algorithm can perform better than the KNN and Naive Bayes algorithm in terms of missed detection rate, false detection rate, sensitivity, and is based on accuracy and specificity. Although the performance of the KNN-NB intrusion detection model is worse than that of the KNN algorithm, it is also within an acceptable range. The accuracy and specificity of the algorithm are better than that of the Naive Bayes algorithm, and its time complexity is better than that of the Naive Bayes algorithm. In general, the intrusion detection model based on KNN-NB algorithm is balanced in various performances. The intrusion detection model based on KNN-NB algorithm is a kind of misuse detection. It has poor detection ability for unknown attack types. Future work can extend it to the detection of unknown attacks.

References

1. Tan, Zhiyuan, et al. Enhancing Big Data Security with Collaborative Intrusion Detection. *IEEE Cloud Computing*, (2015), **1**(3):27-33.
2. Mylavarapu, Goutam, J. Thomas, and T. K. Ashwin Kumar. Real-Time Hybrid Intrusion Detection System Using Apache Storm. *IEEE, International Conference on High PERFORMANCE Computing and Communications IEEE*, (2015):1436-1441.
3. Kietz, J U, et al. "Semantics Inside!" But let's not tell the Data Miners: Intelligent Support for Data Mining. *The Semantic Web: Trends and Challenges. Springer International Publishing*, (2014):706-720.
4. Kulariya, Manish, et al. Performance analysis of network intrusion detection schemes using Apache Spark. *International Conference on Communication and Signal Processing IEEE*,(2016):1973-1977.
5. Neethu, B. Classification of Intrusion Detection Dataset using machine learning Approaches. *International Journal of Electronics & Computer Science Engineering*, (2012), **1**(3):1044-1051.
6. Chauhan, Himadri, et al. A Comparative Study of Classification Techniques for Intrusion Detection. *International Symposium on Computational and Business Intelligence IEEE Computer Society*, (2013):40-43.
7. Hua, Hui You, et al. Hybrid Kmeans with KNN for Network Intrusion Detection Algorithm. *Computer Science*, (2016).
8. Si, Haiyang, et al. The Performance Evaluation of Intrusion Detection Evaluation Method Based on Bayesian Theory. *International Conference on*

Wireless Communications, NETWORKING and Mobile ComputingIEEE, (2008):1-4.

9. Yao, Wei, J. Wang, and S. Zhang. Intrusion detection model based on decision tree and Naive-Bayes classification. *Journal of Computer Applications*, (2015),**7**(12):2883-2885.
10. Quan, Liang Liang, and W. U. Wei-Dong. Anomaly detection model based on support vector machine and Bayesian classification. *Journal of Computer Applications*, (2012), **32**(6):1632-1635.
11. Tavallaee M, Bagheri E, Lu W, et al. A detailed analysis of the KDD CUP 99 data set. *IEEE International Conference on Computational Intelligence for Security and Defense Applications. IEEE Press*, (2009):53-58.
12. Paulauskas N, Auskalnis J. Analysis of data pre-processing influence on intrusion detection using NSL-KDD dataset. *Electrical, Electronic and Information Sciences. IEEE*, (2017):1-5.
13. Alhomoud A, Munir R, Disso J P, et al. Performance Evaluation Study of Intrusion Detection Systems. *Procedia Computer Science*, (2011), **5**(9):173-180.
14. Panigrahi A, Patra M R. Performance Evaluation of Rule Learning Classifiers in Anomaly Based Intrusion Detection. *Computational Intelligence in Data Mining-Volume 2. Springer India*, (2016).
15. Belavagi M C, Muniyal B. Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection. *Procedia Computer Science*, (2016), **89**:117-123.
16. Patel A, Qassim Q, Wills C. A survey of intrusion detection and prevention systems. *Journal of Network & Computer Applications*, (2015), **36**(1):25-41.