A Detailed Analysis of the KDD CUP 99 Data Set

O'rinov Nodirbek Toxirjonovich,
Teacher, Department of Information Technology, Andijan State University
E-mail:nodirbekurinov1@gmail.com

Foziljonova Marxabo
Master of Computer Science and Programming Technology, Andijan State University

ABSTRACT

During in last decade, anomaly detection It has attracted the attention of many researchers to overcome the weakness of signature-based IDS in detecting new attacks, and KDDCUP'99 is the most widely used dataset for grade from these systems. Having held a statistical analysis of this data set, we found two important issues that greatly affects the performance of the systems being evaluated, and the resultsin a very poor evaluation of anomaly detection approaches. To to solve these problems, we proposed a new dataset, NSL-KDD, which the consists from selected records from in full KDD datainstalled as well as does No suffer from Any from mentioned limitations

Keywords:

I. Introduction

With the huge growth in the use of computer networks and the huge increase in the number of applications running on top of them, network security has become increasingly important. As shown in [1], all computer systems have security vulnerabilities, the elimination of which is technically difficult and costly for manufacturers. Therefore, the role of intrusion detection systems (IDS) as specialized devices for detecting anomalies and attacks in the network is becoming increasingly important. Research in the field of intrusion detection has for a long time mainly focused on detection methods based on anomalies and abuses. While commercial products favor misuse-based detection due to its predictability and high accuracy, in academic research, anomaly detection is generally viewed as a more powerful technique due to its theoretical potential to combat new attacks.

By doing a thorough analysis of the recent trend in anomaly detection research, one comes across several machine learning techniques that are reported to have a very high detection rate of 98% while maintaining a false positive rate of 1% [2]. However, when we look at current IDS solutions and commercial tools, we see that there are few products using anomaly detection approaches, and practitioners still think that this is not a mature technology. To find the reason for this contrast, we examined the details of the study done in the field of anomaly detection and looked at various aspects such as training and detection approaches, training datasets, test datasets, and estimation methods. Our study shows that there are some problems in the KDDCUP'99 [3] dataset, which is widely used as one of the few

publicly available datasets for network anomaly detection systems.

The first major disadvantage of the KDD dataset is the huge number of redundant entries. Analyzing the training and test sets of KDD, we found that about 78% and 75% of records are duplicated in the training and test sets, respectively. This large number of redundant entries in the train set will cause the learning algorithms to be biased towards more frequent entries and thus prevent it from learning infrequent entries that are usually more harmful to networks such as U2R attacks. On the other hand, having these duplicate entries in the test set will cause the estimation results to be biased due to methods that have better rates of detecting frequent entries.

In addition, to analyze the complexity level of the records in the KDD dataset, we used 21 trainable machines (7 trainees, each trained 3 times with different sets of trains) to label the records of the entire trainset and the KDD tests, which gives us with 21 predicted labels for each entry. Surprisingly, about 98% of the records in the training set and 86% of the records in the test set were correctly classified by all 21 students. The reason we got these statistics for both the KDD train set and the test sets is because many articles use random parts of the KDD train set as test sets. As a result, they achieve a classification rate of around 98% by applying very simple machine learning techniques. Even applying the KDD test suite will result in a minimum classification rate of 86%, which makes comparing IDS quite difficult as they all range from 86% to 100%.

AT this is paper, we have on condition a solution to decide in the two problems mentioned leading to new sets of trains and tests which consist of selected records of complete KDD data installed. The provided dataset does not suffer from any of the mentioned Problems. Furthermore, in amount from recordsin in train as well as test sets are reasonable. This advantage allows for full experimentation. installed without in need to by chance Choose a small part. Therefore, the evaluation of the results of various studies will to be consistent as well as comparable.

A new version of the KDD dataset, NSL-KDD is publicly available. available to researchers on our website. Although data installed Still suffering from a little from in Problems discussedMcHugh [4] and may not be an ideal representative existing real networks due to lack of publicly available datasets for network IDS, we believe that it can still be applied as effective benchmark dataset to help researchers compare another invasion detection methods.

The rest of the test is organized as follows. Section II represents the KDDCUP99 data set, which is widely used in anomaly detection. AT Chapter III, we the first review in questions in DARPA'98 and then discuss the possible existence these problems in KDD'99. Statistical observations for in KDD data installed will to be explained in Chapter IV. Chapter V offers some solutions to existing problems in KDD data installed. Finally, in Chapter VI we paint conclusion.

II. Kdd A Cup 99 Data Installed Description

Since 1999, KDD'99 [3] has become the most widely used dataset for evaluating anomaly detection methods. This dataset was prepared by Stolfo et al. [5] and built on the basis of data obtained in the IDS DARPA'98 evaluation program [6]. DARPA'98 is about 4 gigabytes of compressed raw (binary) tcpdump data of network traffic over 7 weeks, which can be converted into about 5 million connection records, each about 100 bytes in size. There were about 2 million connection records in two weeks of test data. The KDD training dataset consists of approximately 4,900,000 separate connection vectors, each containing 41 features and labeled as normal or as an attack, with only one specific type of attack. Simulated attacks fall into one of the following four categories:

1) **Negation from service Attack (denial of service):** is en attack in which the attacker does some computation or memory the resource is too busy or full to process legitimate requests, or denies law users access to a car.
2) **User to Root Attack (U2R):** an exploit class in which the attacker starts by accessing the normal user account on the system (perhaps obtained by

listening passwords, dictionary attack or social engineering) and can use some vulnerabilities to get root access to in system.

3) **Remote to Local Attack (R2L):** occurs when en attacking who It has in ability to Send packages to a computer on the network, but which does not have account on this machine uses some vulnerability togrowth local access as a user from what car.
4) **Probing attack:** an attempt to gather information about a network of computers with an explicit purpose from detour this is safety controls.

it is important to note what in test data is No from in same probability distribution as training data, and it includes specific types of attacks that are not in the training data that make task is more realistic. Some invasion experts believe that most new attacks are variants of known attacks, and the signature known attacks may be enough to discover new variants. The data sets contain a total of 24 training attacks. types, with an additional 14 types in the test data only. name and detailed description of the types of training attacks listed in [7].

KDD'99 Peculiarities Can to be secret in three groups:

1) **Base Functions:** this is category encapsulates all in attributes that can be retrieved from a TCP/IP connection. Most of these signs leading to implicit detain in detection.
2) **Traffic features:** this category includes features that are calculated With respect to a window interval as well as is divided in two groups:
   a) **Features of "Same Host":** check only connections in the last 2 seconds that have the same destination host as the current connection, and calculate statistics related to protocol behavior, service, etc.
   b) **"Same service" Functions:** research Only in connections in the last 2 seconds that have same service as in Current connection.

The above two types of "traffic" characteristics called time based. However, there are several slowprobing attacks that scan hosts (or ports) using for example, a much larger time interval than 2 seconds, one every minute. As a result, these attacks create intrusion patterns with time window 2 seconds. To solve this "same host" problem and The characteristics of the "same service" are recalculated, but based on connection window 100 connections, not time window 2 seconds. These features are called connection based traffic Peculiarities.

3) **Content Features:** Unlike most DoS and Probing attacks, R2L and U2R attacks do not have frequent, sequential intrusion patterns. This is because DoS and Probing attacks involve many connections to some hosts in a very short period of time; however, R2L and U2R attacks are built into the data portions of the packets and usually involve only one connection. In order to detect such attacks, we need some functionality to be able to look for suspicious behavior in a piece of data, such as the number of failed login attempts. These features are called content features.

III. Integral Problems From Kdd'99 Data Installed

As it is mentioned in in previous chapter, KDD'99 is builtbased on data received in DARPA'98, which was criticized by McHugh [4], mainly due to the nature of the synthetic data. As a result, some of the existing problems in DARPA'98 remain in KDD'99. However, there some intentional or unintentional enhancements, along with With additional Problems. AT in next we the first review in questions in DARPA'98 and then discuss the possible existence from those Problems in KDD'99. Finally, we discuss new questions observable in in KDD data installed.

1) Per in sake from Confidentiality, in experiments chose to synthesize both in background as well as in attack data,and claims that the data are similar to those observed over several months of data

sampling from a series air force bases. However, neither analytical nor experimental verification of data false alarm characteristics we undertaken. Furthermore, in loadthe synthesized data does not seem to look like in traffic in real networks. Traffic collectors such as TCP dump that is used in DARPA'98 is likely to be overloaded and drop packages in heavy traffic load. However, there was no examination to check the possibility of a fall packages.

2) There is No accurate definition from in attacks. Per example, probing is not necessarily a type of attack if amount from iterations exceeds en specific threshold. Similarly, a packet that causes a buffer overflow does not always represents an attack. In such circumstances, there should be agreement on definitions between the appraiser and the appraiser. In DARPA'98, however, there are no specific network definitionsattacks.

In addition, there is some criticism of attack taxonomies. as well as performance measures. However, these questions are Nois of great interest for this article, since most of the anomaly detection systems work with binary tags, i.e. anomalous and fine, not the definition of detailed information from in attacks. Except, in performance measure applicablein DARPA'98 evaluation, ROC curves are widely was criticized, and since then many researchers have proposed new measures to overcome existing shortcomings [8], [9], [ten], [eleven], [12].

While McHugh's criticisms were mainly based on the procedure for creating the dataset rather than analysis data, Mahoney and Chan [13] analyzed the DARPA background.network traffic and evidence of modeling artifacts found which can lead to an overestimation of the effectiveness some anomaly detection methods. In their work, the authors mentioned five types from anomalies leading to attack detection. However, analysis of the attacks in the DARPA dataset revealed what a lot of did No fit in Any from these categories which the probably caused by modeling artifacts. For example, TTL (time to live) values 126 and 253 only appear in hostile traffic, while in most cases the background traffic value is 127 as well as 254. In a similar way, a little attacks Can to be identified abnormal source IP addresses or anomalies in the TCP protocol window the size field.

Fortunately, the aforementioned modeling artifacts are notaffect the KDD dataset because the 41 functions used in KDD are not associated with any of the disadvantages mentioned in [13]. However, KDD suffering from additional Problems No existing in in DARPA data installed.

AT [fourteen], Tailor this others divided in KDD data installed in ten subsets, each containing approximately 490,000 instances or ten% from in data. However, They observable what in distributionfrom in attacks in in KDD data installed is very uneven which themade cross validation very difficult. Many of these subsets contains instances of only one type. For example, 4th, 5th, 6th, as well as 7th, ten% portions from in full data installed only contained a *smurf* attacks, and data instances in 8th subset we nearly fully *Neptune* intrusions.

Exactly the same problem with *smurf* and *neptune* attacks in the training dataset, KDD is reported in [15]. The authors have mentioned two Problems caused on including theseattacks in the dataset. First, these two types of DoS attacks make up more than 71% of the testing data set, which is completely affects in grade. Secondly, because the They to generate big

TABLE I
STATISTICS OF EXCESSIVE RECORDS IN KDD TRAIN SET

	Original Records	**Distinct Records**	**decline Evaluate**
attacks	3 925 650	262 178	93.32%

Ordinary	972 781	812 814	16.44%
General	4 898 431	1 074 992	78.05%

TABLE II
STATISTICS OF EXCESSIVE RECORDS IN KDD TEST KIT

	Original Records	**Distinct Records**	**decline Evaluate**
attacks	250 436	29 378	88.26%
Ordinary	60 591	47 911	20.92%
General	311 027	77 289	75.15%

volumes of traffic, they are easily detected by other means and there is no need to use anomaly detection systems to find these attacks.

IV. Statistical Observations

As mentioned earlier, there are some problems in KDD data installed which the cause in grade results on the this is data installed to to be unreliable. AT this is chapter we fulfill a installedfrom experiments to show in existing limitations in KDD.

A. Excess Records

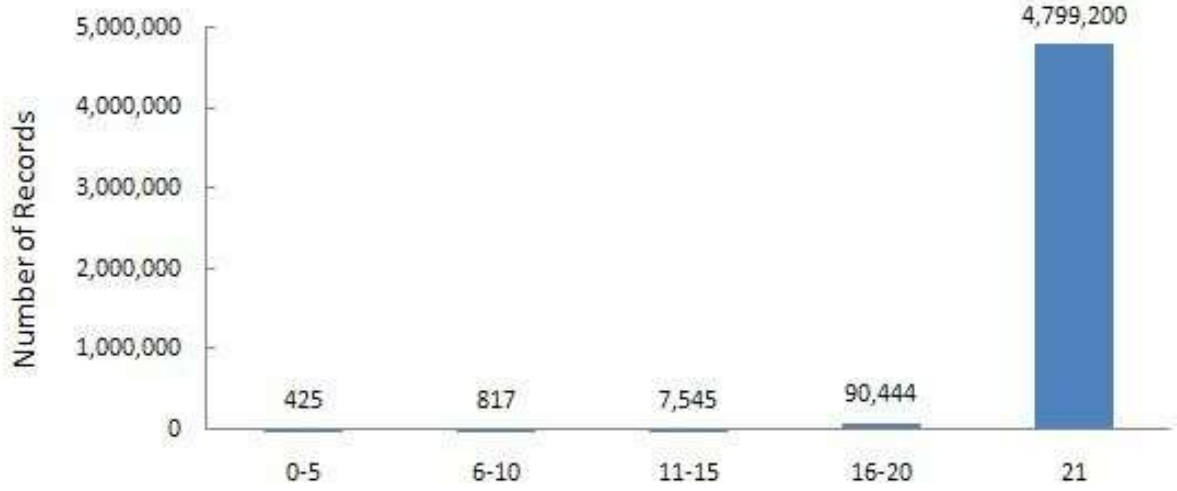
One from in most important limitations in in KDD data installed a huge number of redundant entries, which leads to education algorithms to to be biased in the direction in frequent records,as well as thus prevent them from education infrequent records which the are usually more harmful to networks such as U2Rand R2L attacks. In addition, the presence of these recurring entries in the test set will cause the evaluation results to be biased methods that have better detection rates on in frequent records.

To decide this is problem, we remote all in repeated recordsin the entire KDD train and test set and kept only one copy each entry. Tables I and II illustrate the statistics reducing duplicate entries in the KDD train and test sets, respectively.

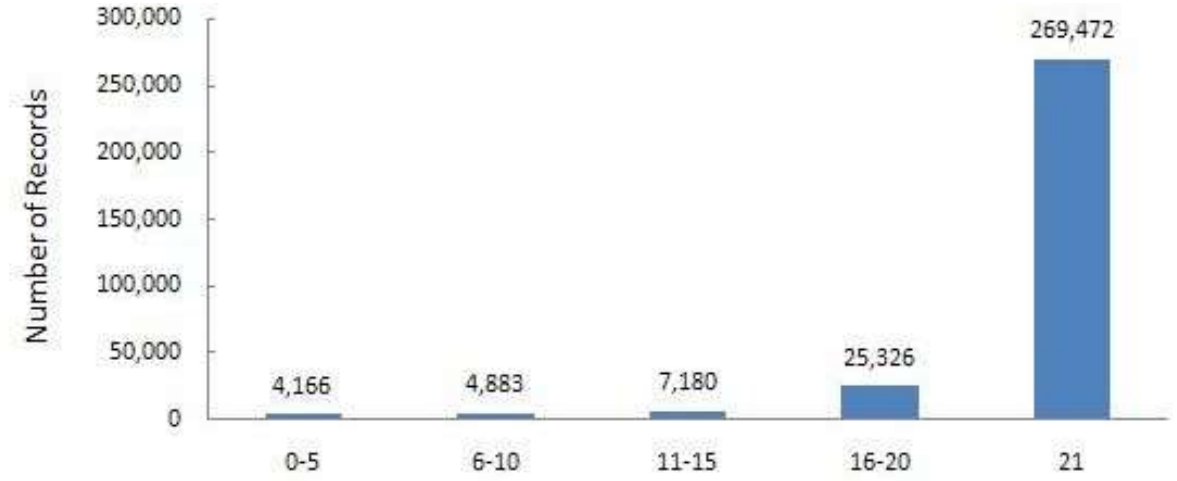
Bye does this is process, we faced two disabled person entries in the KDD test set numbered 136489 and 136497. These two entries contain an invalid ICMP value because they *service* feature. Therefore, we removed them from the KDD test installed.

B. Level from Complexity

Typical Anomaly Detection Approach the use of the KDD data set is to use an individual machine learning algorithm to learn the general behavior of the data installed in order to be able to distinguish between normal and malicious activity. To do this, the dataset is split into test and training segments, where the student is studying using in preparation part from in data installed as well as is then evaluated



Figs. 1. Distribution of #successfulPrediction values for KDD dataset records



Figs. 2. Distribution of #successfulPrediction values for KDD dataset records

for its effectiveness on the test part. Many researchers in in General field from car education have attempt to develop integrated learners to optimize accuracy and detection evaluate above in KDD'99 data installed. AT a similar an approach, we have selected Seven wide used car education technology, namely learning the decision tree J48 [16], naive Bayes [17], NB Wood [eighteen], Random Forest [19], Random Wood [twenty], Multi-layer perceptron [21], as well as Support Vector Car (SVM) [2 2] from in Weka [23] collection to to study in general behavior of the KDD'99 data set. For experiments we applied Weka default values as input parameters these methods.

Examining Existing Anomaly Detection Works which the have used in KDD data installed, we found what there are two general approaches to apply KDD. AT in the first, The training part of KDD'99 is used to sample both training and test sets. However, in the second approach training samples are selected randomly from the KDD train set, and test specimens are randomly selected from in KDD test installed.

For our experiments, we randomly createdthree smaller subsets of the KDD train set, each of which included fifty thousand records of information. Each of the students were trained on the created sets of trains. We then busy in 21 learned cars (7 students, each trained3 times) for marking the records of the whole train and the KDD test sets, which gives us 21 predicate labels for each write down. Farther, we annotated each write down from in data installed With # successful prediction a value that has been initialized to zero. Now, since the KDD dataset provides the correct representation Bel per each write down, we compared in predicative label from each write

down the to a specific student with an actual label where we have increased #successfulForecast by one if a a match has been found. During this process, we calculated amount from students what we able to right label what the write down. The maximum #successfulPrediction value is 21, indicating that all students were able to correctly predict the label of this post. Figures 1 and 2 show the distribution of #successfulPrediction values for the KDD train and test sets, respectively.

it Can to be clearly visible from Figure one as well as 2 what 97.97%as well as 86.64% from in records in in KDD train as well as test setswere correctly labeled by all 21 classifiers. Obvious observation from these numbers is what in statement from

TABLE III
STATISTICS OF RANDOMLY SELECTED RECORDS FROM KDD TRAIN SET

	Distinct Records	**Percent**	**Selected Records**
0-5	407	0.04	407
6-10	768	0.07	767
11-15	6525	0.61	6485
16-20	58 995	5.49	55 757
21	1 008 297	93.80	62 557
General	1 074 992	100.00	125 973

typical education cars to this is data installed would result inhigh accuracy rates. This shows that the evaluation of methods on in the foundation from accuracy, detection evaluate as well as FALSE positive evaluate on the in KDD data installed is No en corresponding option.

V. Our Solution

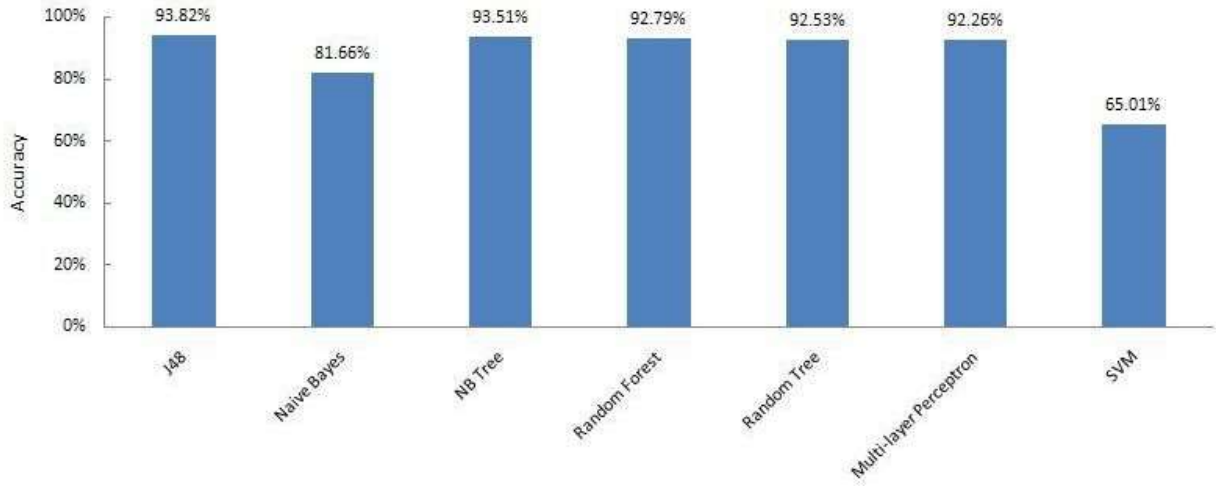
To address the issues mentioned in the previous section, we first removed all redundant entries in both the training and test sets. In addition, to create a more complex subset of the KDD dataset, we randomly selected records from the #successfulPrediction value groups shown in Figures 1 and 2, such that the number of records selected from each group was inversely proportional to the percentage of records in the original groups. #successfulPrediction values . For example, the number of records in the value group 0-5 #successfulPrediction of the KDD trainset is 0.04% of the original records, hence 99.96% of the records in this group are included in the generated sample. Tables III and IV show detailed statistics of randomly selected records.

Generated datasets, KDD Train+ and KDD Test+,

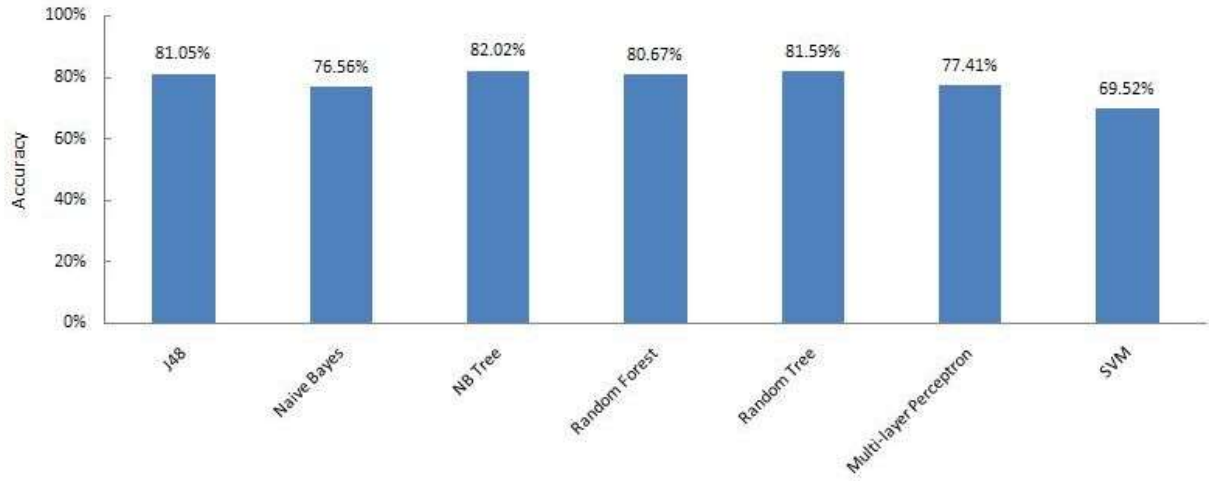
TABLE IV
STATISTICS OF RANDOMLY SELECTED RECORDS FROM KDD TEST KIT

	Distinct Records	**Percent**	**Selected Records**
0-5	589	0.76	585
6-10	847	1.10	838

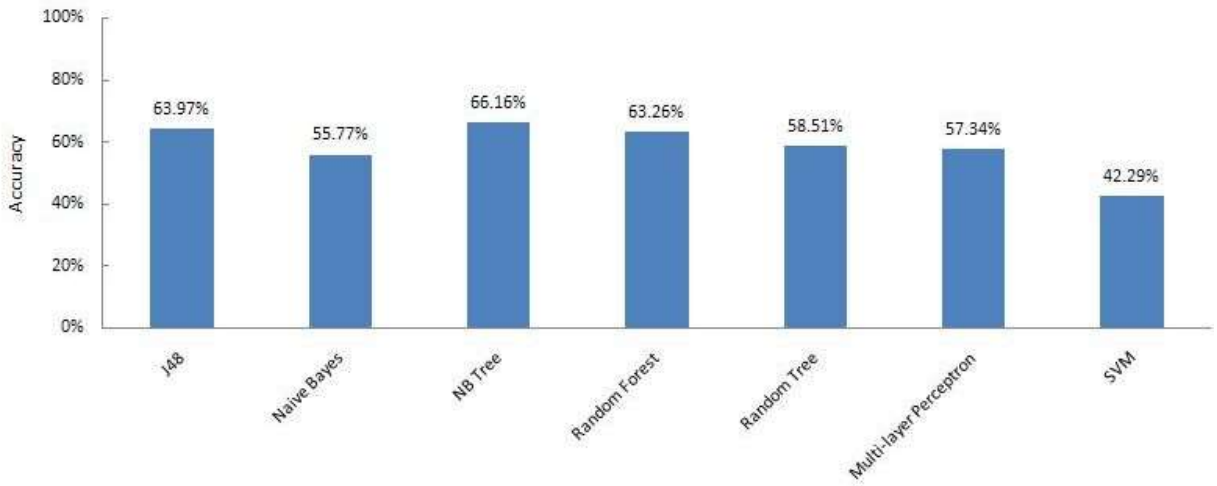
11-15	3540	4.58	3378
16-20	7845	10.15	7049
21	64 468	83.41	10 694
General	77 289	100.00	22 544



Figs. 3. performance from in selected education cars on the KDD test



Figs. four. performance from in selected education cars on the KDD test +



Figs. 5. performance from in selected education cars on the KDD Test -21

included 125,973 and 22,544 entries, respectively. Farther- moreover, another test suite was generated that did not include any of the entries that were correctly classified by all 21 students, KDDTest - 21 , which included 11,850 entries. Per experimental goals, we busy in the first twenty% fromin records in KDD Train + as in train installed, having trained in education methods, we applied in learned models on thethree test suites, namely KDD Test (original KDD test suite), KDD test + as well as KDDTest - 21 . result from in grade students on these datasets are shown in Figures 3, 4. as well as 5, respectively.

As Can to be visible in Figure 3, in accuracy evaluate from in classifiers on the KDD test is relatively high. This shows what the original KDD test set is skewed and disproportionate distributed, making it unsuitable for testing network based on anomaly detection classifiers. The results of the accuracy and performance of learning machines at KDD'99 the data set is therefore unreliable and cannot be used as good indicators from in ability from in classifier to service as a discrimination tool for detecting anomalies in the network. On the on the contrary, the test suites KDD Test + and KDDTest - 21 provide more accurate information about the capabilities of classifiers. As en example, classification from SVM on the KDD testis 65.01% which the is enough poor compared to Another education fits. However, SVM is the only learning method. whose performance is improved on KDD Test + . Analysis both sets of tests, we found that SVM erroneously determines one of the most frequent entries in the KDD test, which greatly affects it detection performance. On the contrary, in KDD Test + , since it is recording occurs only once, it does not affect in classification evaluate from svm, as well as provides better gradefrom education methods

VI. Concluding Remarks

AT this is paper, we statistically analyzed in the whole KDD data set. The analysis showed that there are two important questions in in data installed which the very affects in performanceevaluated systems and leads to a very poor evaluation anomaly detection approaches. To solve these problems, we have proposed a new dataset, NSL-KDD [24], which consists of selected records of the complete KDD dataset. This dataset publicly available to researchers through our website and It has in the following benefits above original KDD data installed:

- It does not include redundant entries in the train set, so classifiers will not be biased towards more frequent records.
- entries in the proposed test suites ; hence student outcomes are not biased methods that have the best detection rates on frequent records.
- The number of selected records from each difficulty - level Group is back proportional to in percentrecords in the original KDD dataset. As a result classification speed of various machine learning methods differ in a Shire

range, which the does it more effectiveto have en accurate grade from another education methods.

- The number of records in train and test sets is reasonable, which makes experimentation accessible on the in full installed without in need to by chance choose a small portion. Therefore, the evaluation results various research papers will be consistent and parable.

Although, in proposed data installed Still suffering from a littlefrom in Problems discussed on McHugh [four] as well as May No to bea perfect representative from existing real networks, becauselack of public datasets for network IDS, we believe that it can still be used as an effective benchmark installed to help researchers compare different intrusion detection systems methods.

References

1. CE Landwehr, AR Bull, JP McDermott and WS Choi, "A classification of security flaws in computer programs", *ACM Comput . Survive _* , about. 26, no. 3, page 211–254, 1994.
2. M. Shew, FROM. Chen TO. Sarinnapakorn , as well as L. chang, "BUT novel anomaly detection scheme based on principal components classifier", *Proceedings of the IEEE Foundations and New Directions of Data Mining Seminar, in association with Third IEEE International Conference on the Data Mining (ICDM03)* , page 172–179, 2003.
3. KDD 1999. Available at: http://kdd.ics.uci.edu/databases/kddcup.99/kddcup99.html, October 2007
4. J. McHugh "Testing invasion detection systems: a criticism from in 1998as well as 1999 Darpa invasion detection system estimates as completed Lincoln Lab, *ACM Transactions on Information and System safety* , about. 3, no. four, page 262–294, 2000.
5. SJ Stolfo, W. Fan, W. Lee, A. Prodromidis and PK Chan , "Cost - based modeling per fraud as well as invasion detection: Results from in jam project," *discex* , about. 02, P. 1130 2000.
6. R. P. Lippmann, D. J. fried, I. Graph, J. AT. haynes, TO. R. Kendall, D. McClung, D. Weber, S. E. Webster, D. Wyshogrod , R. C. Cunningham, as well as M. BUT. Zissman , "Grade invasion detection systems: 1998 darpa offline invasion detection grade," *discex* , about. 02, P. 1012, 2000.
7. MIT Lincoln Labs, 1998 DARPA DARPA Detection Detection Evaluation. Available on: http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/index.html , February 2008.
8. FROM. Axelsson , " base rate delusion as well as in difficulty from invasion detection", *ACM Transactions on Information and System Security (TISSEC)* , about. 3, no. 3, page 186–205, 2000.
9. J. Gaffney Jr as well as J. Ulvila , "Grade from invasion detectors: BUT decision theory approach", in the *proceedings of the IEEE Symposium on Safety as well as Confidentiality, (S&P)* , page 50–61, 2001.
10. J. Di Crescenzo , A. Ghosh, R. Talpade, "Toward a Theory invasion detection", *Lecture notes in a computer science* , about. 3679,P. 267, 2005.
11. BUT. cardenas, J. Baras , as well as TO. Simon, "BUT framework per in assessment of intrusion detection systems, in *Proceedings of IEEE Symposium on the Safety as well as Confidentiality, (S&P)* , P. fifteen, 2006.
12. GRAM. Gu , P.F ogla , D. Dagon, B. Lee, as well as B. S k or i ć , "Dimension invasion detection possibility: An information-theoretical an approach," in *Proceedings of the ACM Symposium on Information, Computers and Communications Security (ASIACCS06)* , pp. 90–101, ACM New York, New York, USA, 2006.
13. M. Mahoney and P. Chan, "DARPA/Lincoln Analysis 1999 Laboratory Evaluation Data for Network Anomaly Detection", *LEC-*

TOUR NOTES AT A COMPUTER SCIENCE , page 220–238, 2003.

14. L. Tailor, E. Eskin, as well as FROM. Stolfo, "Invasion detection With unmarked data using clustering", *Proceedings of ACM CSS Workshop on Data Mining Applicable to Safety, Philadelphia, Pennsylvania, november* , 2001.
15. K. Leung and K. Leki, "Unsupervised Network Anomaly Detection." intrusion detection using clusters", *Materials 28th Australasian Conference on Computer Science, Volume 38* , pp. 333– 342, 2005.
16. J. Quinlan, *C4.5: Programs per Car Training* . Morgan Kaufmann, 1993.
17. GRAM. John as well as P. langley, "Grade continuous distribution in Bayesian classifiers", in the *materials of the eleventh conference on Uncertainty in Artificial intelligence* , page 338–345, 1995.
18. R. Kochavi, "Improving the Accuracy of Naive Bayes Classifiers: A decision tree hybrid," in *Proceedings from in Second International Conference on the Knowledge Opening as well as Data Mining* , about. 7, 1996.
19. L. Breiman , "Random The woods", *Car training* , about. 45, no. one, page 5–32 2001.
20. Aldous, " continuum random wood. I," *Annals from Probability* _ page 1–28 1991.
21. Crayfish, FROM. Rogers, M. Kabriski , M. oxley, as well as B. Suter, " Multilayer Perceptron as an Approximation to the Optimal Bayesian Discriminant Function", *IEEE Transactions on Neural Networks* , vol. 1, no. four, page 296–298, 1990.
22. K. Chang and K. Lin, "LIBSVM: A Library for Support Vector Machines", 2001. Software. accessible in http://www.csie.ntu.edu.tw/cjlin/libsvm .
23. "Waikato Knowledge Analysis Environment (weka) Version 3.5.7". Accessible on the: http://www.cs.waikato.ac.nz/ml/weka /, June, 2008.
24. " Nsl-cdd data installed per network invasion detection systems." Accessible on the: http://nsl.cs.unb.ca/NSL-KDD/, March 2009.