Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 57 (2015) 842 – 851

Procedia Computer Science

3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)

Analysis of KDD Dataset Attributes - Class wise For Intrusion Detection

Preeti Aggarwal[a,]*, Sudhir Kumar Sharma[b]

[a]*School of Engineering and Technology, Ansal University, Gurgaon-122001, India*
[b]*School of Engineering and Technology, Ansal University, Gurgaon-122001, India*

Abstract

The KDD data set is a well known benchmark in the research of Intrusion Detection techniques. A lot of work is going on for the improvement of intrusion detection strategies while the research on the data used for training and testing the detection model is equally of prime concern because better data quality can improve offline intrusion detection. This paper presents the analysis of KDD data set with respect to four classes which are Basic, Content, Traffic and Host in which all data attributes can be categorized. The analysis is done with respect to two prominent evaluation metrics, Detection Rate (DR) and False Alarm Rate (FAR) for an Intrusion Detection System (IDS). As a result of this empirical analysis on the data set, the contribution of each of four classes of attributes on DR and FAR is shown which can help enhance the suitability of data set to achieve maximum DR with minimum FAR.

Peer-review under responsibility of organizing committee of the 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)

Keywords: Intrusion Detection; KDD; Attribute classes; False Alarm Rate; Detection Rate.

1. Introduction

Network security[1] is a foremost issue these days as the network usage is growing in multi-dimensions due to increased use of handheld devices. Intrusion Detection Systems can help detect malign intentions of network users without compromising the security of the host and the network. There are many machine learning algorithms

* Corresponding author. Tel.: +91-987-329-9946.
*E-mail address:*preetaggarwal@gmail.com

Peer-review under responsibility of organizing committee of the 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)
doi:10.1016/j.procs.2015.07.490

available which can learn from the training data and can generalize when exposed to new untrained data. There are two types of intrusion detection technique, the first one is Misuse Detection that can catch the known attacks and hence works on the offline data[2] and the other is Anomaly Detection which can detect any abnormal behavior and hence can work well on online data[3]. The KDD data set is a standard data set used for the research on intrusion detection systems.

1.1. KDD Data Set

The NSL-KDD data set with 42 attributes is used in this empirical study. This data set is an improvement over KDD'99 data set[4, 5] from which duplicate instances were removed to get rid of biased classification results[6-9]. This data set has number of versions available, out of which 20% of the training data is used which is identified as KDDTrain+_20Percent with a total number of 25192 instances. The test data set is identified by the name KDDTest+ and has a total of 22544 instances. Different configurations of this data set are available with variation in number of instances but the number of attributes in each case is 42. The attribute labeled 42 in the data set is the 'class' attribute which indicates whether a given instance is a normal connection instance or an attack. Table 1 gives the description of KDD data set attributes with class labels. Out of these 42 attributes, 41 attributes can be classified into four different classes as discussed below[10]:

- Basic (B) Features are the attributes of individual TCP connections
- Content (C) features are the attributes within a connection suggested by the domain knowledge
- Traffic (T) features are the attributes computed using a two-second time window
- Host (H) features are the attributes designed to assess attacks which last for more than two seconds

Table 1.Classwise detail of KDD data set attributes.

Sr. No	Label	Attribute Name	Sr. No	Label	Attribute Name	Sr. No	Label	Attribute Name	Sr. No	Label	Attribute Name
1	B	duration	10	C	hot	23	T	count	32	H	dst_host_count
2	B	protocol_type	11	C	num_failed_logins	24	T	serror_rate	33	H	dst_host_srv_count
3	B	service	12	C	logged_in	25	T	rerror_rate	34	H	dst_host_same_srv_rate
4	B	src_bytes	13	C	num_compromised	26	T	same_srv_rate	35	H	dst_host_diff_srv_rate
5	B	dst_bytes	14	C	root_shell	27	T	diff_srv_rate	36	H	dst_host_same_src_port_rate
6	B	flag	15	C	su_attempted	28	T	srv_count	37	H	dst_host_srv_diff_host_rate
7	B	land	16	C	num_root	29	T	srv_serror_rate	38	H	dst_host_serror_rate
8	B	wrong_fragment	17	C	num_file_creations	30	T	srv_rerror_rate	39	H	dst_host_srv_serror_rate
9	B	urgent	18	C	num_shells	31	T	srv_diff_host_rate	40	H	dst_host_rerror_rate
			19	C	num_access_files				41	H	dst_host_srv_rerror_rate
			20	C	num_outbound_cmds				42	-	class
			21	C	is_hot_login						
			22	C	is_guest_login						

1.2. Objective

The objective of this research work is to study the NSL-KDD data set[10] from the viewpoint of four classes of attributes rather than study the behavior of individual attribute. The four classes of attributes are Basic, Content, Traffic and Host. For evaluation of Intrusion Detection, there are number of metrics available but in this work, two prominently used metrics are discussed, i.e., FAR and DR[11]. This paper presents the contribution of each of four classes of attributes in the evaluation of FAR and DR metrics. Hence in the later part of the paper, it is concluded how each attribute class impacts the two metrics.

1.3. Related Work

The intrinsic problem of KDD data set has led to the development of new data set known as NSL-KDD data set[10]. This new data set has overcome many problems like redundant instances[6, 7]. The problem of redundancy results in biased results, that is, if a certain instance is repeated many times, the leaning gets biased. This is one of the reasons why certain classifiers show accuracy of above 95% as well[8, 9] for intrusion detection. The study shows that the machine learning algorithms does not produce good results in case of misuse detection[2]. Various algorithms are evaluated for IDS[12] which involves rule based classifiers[13] and decision theory approach[14]. In the Weka tool[15], many variants of basic classifiers are implemented which generate quite interesting results. Some of the variants of tree based algorithms are random tree[16-18] and random forest algorithm.

The rest of the paper is organized as follows: Section 2 presents the research methodology involving the tool, used classifier and the evaluation metrics. Section 3 presents the simulation results with section 4 discussing the results. Section 5 presents the conclusion of this empirical study.

2. Experimental Setup

2.1. Research Methodology

The steps followed as part of the research methodology are as follows:

- KDD data set is selected[10]
- Weka Tool is chosen for simulation
- Random Tree is used as a binary classifier for simulation on Weka classifies the instances as attack or normal
- Preprocessing of training and testing data file with 42 attributes is done to generate 14 new training data files for each combination as discussed in Table 2
- Every pair of 15 data set files (training and test), is simulated on random tree algorithm and the results are tabulated in Table 6

It must be noted that all 15 training and test data files as in Table 2, the last attribute of the original data set, that is, ‘class’ attribute is included.

2.2. Weka

Waikato Environment for Knowledge Analysis (Weka)[15] is a data mining tool available free of cost under the GNU General Public License. The version used in this study is 3.7.11 that has many state of the art machine learning tools and algorithms for data analysis and predictive modeling. This tool accepts the data file either in comma separated value (csv) or attribute-relation file format (arff) file format. For the simulation, arff files is already available with 42 attributes whereas arff files with lesser attributes as discussed in research methodology section are created through the pre-processing tab of the tool.

Table 2.Combinations of attribute classes for KDD data set.

Sr. No.	Attribute class Combinations	# Attributes	B	C	T	H
1	BCTH	41	√	√	√	√
2	BCT	31	√	√	√	x
3	BCH	32	√	√	x	√
4	BTH	28	√	x	√	√
5	CTH	32	x	√	√	√
6	BC	22	√	√	x	x
7	BT	18	√	x	√	x
8	BH	19	√	x	x	√
8	CT	22	x	√	√	x
10	CH	23	x	√	x	√
11	TH	19	x	x	√	√
12	B	9	√	x	x	x
13	C	13	x	√	x	x
14	T	9	x	x	√	x
15	H	10	x	x	x	√

2.3. Used Classifier

Machine learning[12] is an artificial intelligence technique which consists of a number of algorithms based on which a model can be developed that learns from the input data known as the training data set and helps predict on testing data set[13]. Though there are many classifiers available[12], but tree based algorithms[13] produce better accuracy in results without requiring much tuning of parameters. In this paper, Random Tree algorithm, a tree based classifier[14, 19, 20] is selected for simulation from past experience. Random Tree[16, 17] is a set (ensemble) of tree predictors that is called forest. This classifier takes the input feature vector, classifies it with every tree in the forest, and outputs the class label that receives the majority of "votes"[16].

2.4. Metrics

Intrusion detection metrics helps evaluate the performance of an intrusion detection system[21]. Some of the commonly used evaluation metrics used with respect to intrusion detection are False Alarm Rate (FAR), Detection Rate (DR), Accuracy, Precision, Specificity, F-score[13].

All these evaluation metrics are basically derived from the four basic attributes of the confusion matrix depicting the actual and predicted classes. These elements of the confusion matrix are:

- True Negative (TN): Number of instances correctly predicted as non-attacks.
- False Negative (FN): Number of instances wrongly predicted as non-attacks.
- False Positive (FP): Number of instances wrongly predicted as attacks.
- True Positive (TP): Number of instances correctly predicted as attacks.

As shown in the Table 6, all the metrics are generated from these four basic elements. In this paper, two of the evaluation metrics that are considered for this study are FAR which is defined as the rate at which normal instances are classified as anomalous and DR which is defined as the ratio of number of instances of correctly predicted attacks to the total number of actual attack instances. Another metric used for the analysis of results is the graphical plot known as the Receiver Operating Characteristic (ROC) curve. It is a plot of DR and FAR. Though this curve does not exactly tell the best classification results in terms of FAR and DR but the area under the ROC curve helps

decide the best possible combination of DR and FAR because the best intrusion detection system is one with maximum DR and minimum FAR at the same time.

3. Simulation Results

The classification results produced by Weka tool are in the form of a confusion matrix which gives the actual versus predicted classification results. The results generated for fifteen data files and the calculated evaluation metrics are tabulated in Table 6 (Appendix). It can be observed that only the frequently used metrics are tabulated out of which, the focus of study is on DR and FAR. Fig. 1 below shows the bar graph presenting DR and FAR for each of fifteen sets of data. Fig. 2 depicts the ROC curve with single class of attributes, and the curve also shows the plot of all the 41 attributes (four classes together). Similarly, Fig. 3 and Fig. 4 show the ROC curve with 2 classes of attributes and three classes of attributes respectively. The arrow on the ROC curve shows the best possible class of data sets with maximum DR and minimum FAR. Key observations from ROC curve of Fig. 2 are as follows:

- Basic class attributes show higher DR
- Traffic attributes show lower FAR
- Content attributes show higher FAR
- Host attributes show low DR but decent FAR

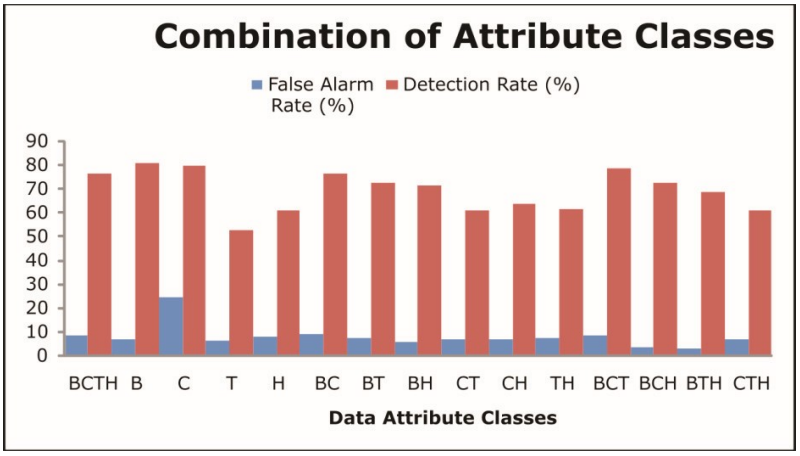


Fig. 1. Detection rate and False alarm rate for 15 combinations of attribute classes.

Key observations from ROC curve of Fig. 3 are as follows:

- BC class attributes show DR equivalent to BCTH but has higher FAR
- Without basic class attributes, DR drops

Key observations from ROC curve of Fig. 4 are as follows:

- With Traffic class attributes, FAR is comparatively higher
- Presence of Basic class attributes show higher DR

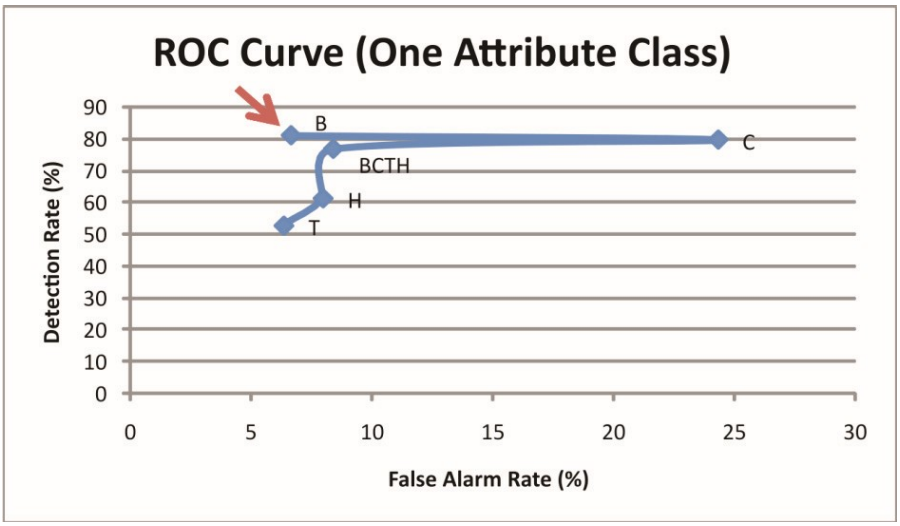


Fig. 2. ROC curve for single class of attributes.

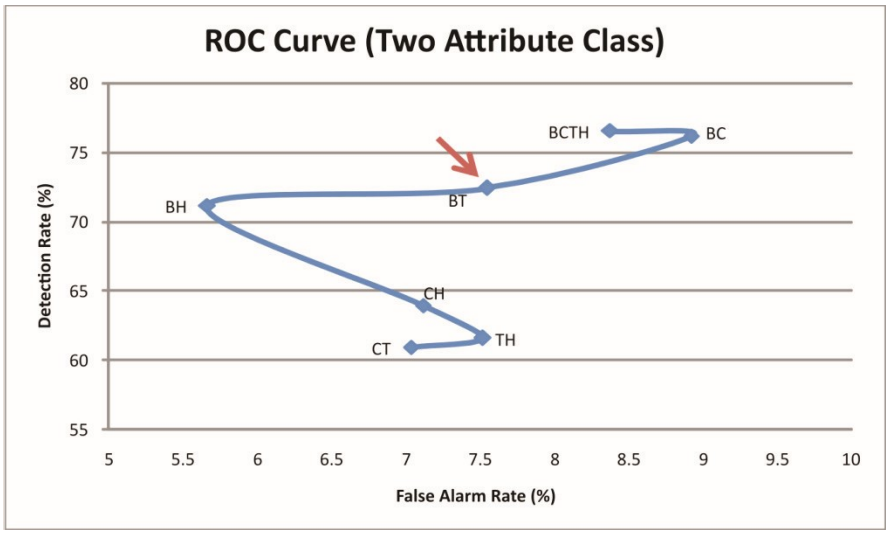


Fig. 3. ROC Curve for two classes of attributes.

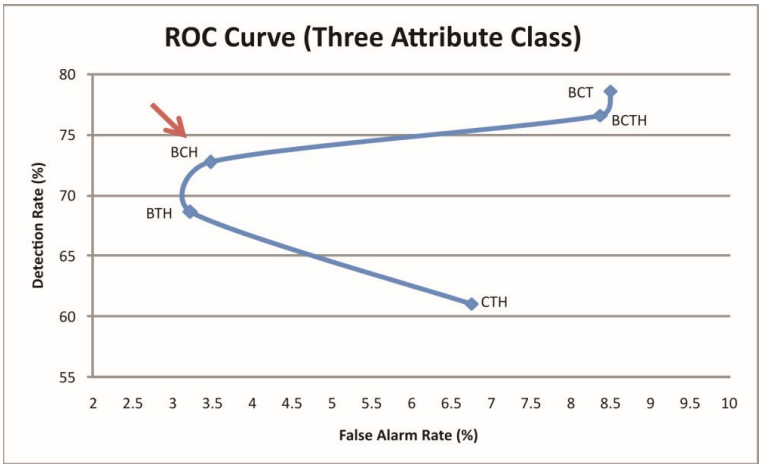


Fig. 4. ROC Curve for three classes of attributes.

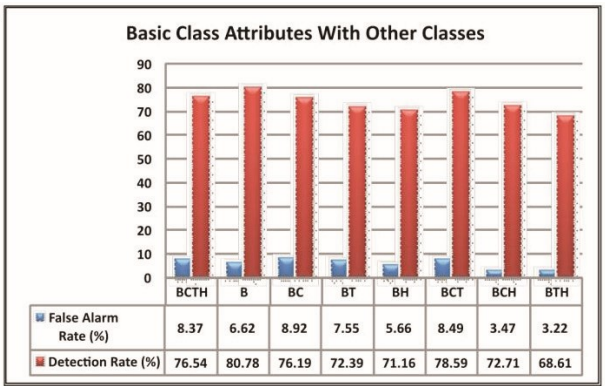
Further analysis of DR and FAR is done with respect to each class of attributes. For this, Table 6 is referred and those entries are selected where the one of the four classes under observation is definitely present. Following this analysis pattern, Fig. 5(a) shows all those attribute class's combination where basic class is definitely present. Therefore, this plot shows the contribution of the attributes of the basic class prominently. Similarly, Fig. 5 to 6 show the combinations of those attribute classes where content, traffic and host attributes respectively are definitely present. Key observations from Fig. 5(a) where basic class attributes are present in combinations with other class attributes are as follows:

- DR is above 70 % for almost all the combinations
- Traffic attributes show lower FAR

Key observations from Fig. 5 (b) where content class attributes are present in combinations with other class attributes are as follows:

- DR is highest for content class attributes
- DR improves wherever basic class attributes are involved
- FAR is maximum for content class attributes

a



b

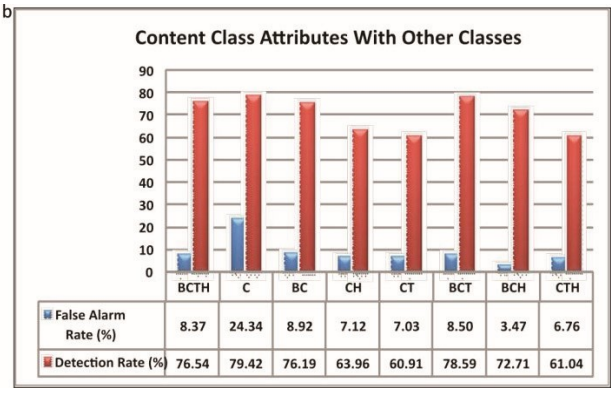
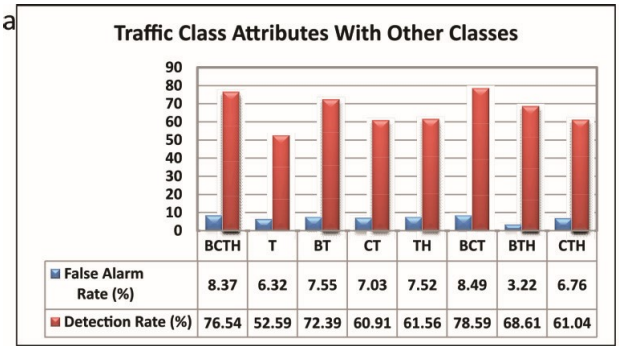


Fig. 5. (a) Plot of DR and FAR for attributes of basic class with other attribute classes; (b) Plot of DR and FAR for attributes of content class with other attribute classes.

a



b

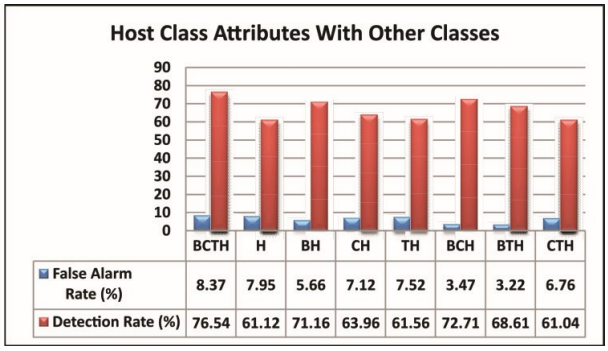


Fig. 6. (a) Plot of DR and FAR for attributes of traffic class with other attribute classes; (b) Plot of DR and FAR for attributes of host class with other attribute classes.

Key observations from Fig. 6 (a) where traffic class attributes are present in combinations with other class attributes are as follows:

- DR is minimum with Traffic class attributes alone
- DR improves wherever basic class attributes are involved

Key observations from Fig. 6(b) where host class attributes are present in combination with other class attributes is that lower FAR is observed with basic and host class attribute combinations.

4. Discussions

Table 3 and 4 are generated in reference to Fig. 5 and Fig. 6. Table 3 presents the best values of DR (maximum) and FAR (minimum) with respect to each class of attributes in combination with other attribute classes. Similarly, Table 4 presents the worst values of DR (minimum) and FAR (maximum) such that the first row shows the worst values for all the attribute class' combinations where Basic class is definitely present. Second row show the values when Content attributes are definitely present. It can be observed from Table 3 and 4 that the presence of basic class attributes show maximum DR and the traffic class attributes show lower DR. Similarly Table 3 shows that FAR is more when content class attributes are definitely present but consistent otherwise.

Table 3.Best Detection and False Alarm Rate for each class of attributes.

Attribute Class	Best Detection Rate (%)	Best False Alarm Rate (%)
Basic	**80.78**	**3.22**
Content	79.42	3.47
Traffic	78.59	**3.22**
Host	76.54	**3.22**

Table 4.Worst Detection and False Alarm Rate for each class of attributes.

Attribute Class	Worst Detection Rate (%)	Worst False Alarm Rate (%)
Basic	68.61	8.92
Content	60.91	**24.34**
Traffic	**52.59**	8.50
Host	61.04	**8.37**

Table 4 also shows a similar scenario of worst FAR in the case of content class attributes. The other observation of critical importance is that the host class attributes show best FAR value even in the worst case. Table 5 presents the overall result summary of this empirical study.

Table 5.Overall result summary.

Attribute Class	Prevailing Contribution
Basic	High DR
Content	High FAR
Traffic	Low DR
Host	Low FAR

5. Conclusion

This paper used four categories of attributes Basic, Content, Traffic and Host in which 41 attributes of KDD data set were categorized and fifteen variants of data set were generated by forming all combinations of four classes. These fifteen sets of training and test data files were simulated on Random Tree algorithm in Weka tool. The results were analyzed to study dominance of each class of attributes in improving the Detection Rate (DR) and minimizing the False Alarm Rate (FAR). This study can help increase the suitability of the data set so that higher DR can be achieved with minimum FAR. Hence, future work can lead to an improved data set that can be utilized for online intrusion detection.

Appendix A.

The appendix presents the summary of result which is generated using Weka tool for Random Tree algorithm on fifteen data set configurations of KDD[4] data set.

Table 6.Summary result for Random Tree algorithm.

Sr. No.	#Attribute Classes	Attribute Class Combinations	TN	FN	FP	TP	Accuracy	False Alarm Rate	Detection Rate	F-Score
1	4	BCTH(41)	8898	3011	813	9822	83.04	0.08	0.77	0.84
2	3	BCT(31)	8886	2748	825	10085	84.15	0.08	0.79	0.85
3	3	BCH(32)	9374	3502	337	9331	82.97	0.03	0.73	0.83
4	3	BTH(28)	9398	4028	313	8805	80.74	0.03	0.69	0.80
5	3	CTH(32)	9055	5000	656	7833	74.91	0.07	0.61	0.73
6	2	BC(22)	8845	3056	866	9777	82.6	0.09	0.76	0.83
7	2	BT(18)	8978	3543	733	9290	81.03	0.08	0.72	0.81
8	2	BH(19)	9161	3701	550	9132	81.14	0.06	0.71	0.81
9	2	CT(22)	9028	5016	683	7817	74.72	0.07	0.61	0.73
10	2	CH(23)	9020	4625	691	8208	76.42	0.07	0.64	0.76
11	2	TH(19)	8981	4933	730	7900	74.88	0.08	0.62	0.74
12	1	B(9)	9068	2466	643	10367	86.21	0.07	0.81	0.87
13	1	C(13)	7347	2641	2364	10192	77.79	0.24	0.79	0.80
14	1	T(9)	9097	6084	614	6749	70.29	0.06	0.53	0.67
15	1	H(10)	8939	4989	772	7844	74.45	0.08	0.61	0.73

References

1. Kumar, Vipin, Jaideep Srivastava, and Aleksandar Lazarevic, "Managing cyber threats: Issues, approaches, and challenges". Vol. 5.Springer, 2006.
2. Maheshkumar Sabhnani and Gursel Serpen, "Why Machine Learning Algorithms Fail in Misuse Detection on KDD Intrusion Detection Data Set". ACM Transactions on Intelligent Data Analysis,(pp.403-415) (2004).
3. M. Shyu, S. Chen, K. Sarinnapakorn, and L. Chang, "A novel anomaly detection scheme based on principal component classifier". Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with the Third IEEE International Conference on Data Mining (ICDM03), pp. 172– 179, 2003.
4. KDD Cup 1999. (2014, Nov.) [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/
5. MIT Lincoln Labs. (2014, Nov.). DARPA intrusion detection evaluation [Online]. Available: http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/index.html
6. J. McHugh, "Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory". ACM Transactions on Information and System Security, vol. 3, no. 4, pp. 262–294, 2000.
7. Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani, "A Detailed analysis of the KDD CUP 99 Data Set". In the Proc. Of the IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009), pp. 1-6, 2009.
8. S. Revathi, Dr. A. Malathi, "A detailed analysis of KDD cup99 Dataset for IDS". International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 12, December – 2013.
9. R. P. Lippmann, D. J. Fried, and I. Graf, "Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation". In Proceedings of the 2000 DARPA Information Survivability Conference and Exposition (DISCEX'00), (2000).
10. "Nsl-kdd data set for network-based intrusion detection systems". Available on: http://nsl.cs.unb.ca/NSL-KDD/, November 2014.
11. Lee W., and Stolfo S.J., "A framework for constructing features and models for intrusion detection systems". ACM Transactions on Information and System Security, 3 (4) (pp. 227-261) (2000).
12. Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg, " Top Ten Data Mining Algorithms". Knowledge and Information Systems Journal, Springer-Verlag London, vol. 14, Issue 1, pp. 1-37, 2007.
13. Lei Li, De-Zhang Yang, Fang-Cheng Shen, "A Novel Rule-based Intrusion detection System Using Data Mining". In the Proc. Of 3rd IEEE International Conference on Computer Science and Information Technology, pp. 169-172, 2010.
14. J. E. Gaffney and J. W. Ulvila, "Evaluation of intrusion detectors: A decision theory approach". In Proceedings of the 2001 IEEE symposium on Security and Privacy, pages 5061, Oakland, CA, USA, 2001.
15. "Waikato environment for knowledge analysis (weka) version 3.7.11." Available on: http://www.cs.waikato.ac.nz/ml/weka/, November, 2014.
16. D. Aldous, "The continuum random tree. I". The Annals of Probability, pp. 1–28, 1991.
17. Breiman, Leo, Friedman, J. H., Olshen, R. A., Stone, C. J., "Classification and regression trees". Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.ISBN 978-0-412-04841-8. (1984)
18. Data Mining Practical Machine Learning Tools and Techniques by Ian H Witten, Eibe Frank, Mark A Hall.
19. Han, Jiawei, and Micheline Kamber. Data Mining, Southeast Asia Edition: Concepts and Techniques. Morgan kaufmann, 2006.
20. Preeti Aggarwal, Sudhir K Sharma, "An Empirical Comparison of Classifiers to Analyze Intrusion Detection". International Conference on Advanced Computing & Communication Technologies, IEEE Feb-2015.
21. G. Gu, P. Fogla, D. Dagon and W. Lee, "An Information-Theoretic Measure of Intrusion Detection Capability". In Proceedings of the 2006 ACM Symposium on Information, computer and communications security; 21-24 Mar. (2006).