The Effect of Normalization on Intrusion Detection Classifiers (Naïve Bayes and J48)

D. Ashok Kumar
Associate Professor,
Department of Computer Science,
Government Arts College
Tiruchirappalli, India
e-mail: akudaiyar@yahoo.com

S. R. Venugopalan
Scientist.
Information & Computing Technologies Directorate
Aeronautical Development Agency
(Ministry of Defence; Govt. of India), Bangalore, India
e-mail:venu_srv@yahoo.com

Abstract—Intrusion Detection has become an inevitable area for commercial applications and academic research. Network traffic is typically very high volume and consists of both qualitative and quantitative data with different range of values. Raw data needs to be pre-processed before fed into any learning model and the most used technique is normalization [1]. Attribute normalization eliminates the dominance of attributes with extreme values by scaling it within the range. However, many intrusion detection methods do not normalize attributes before training and detection [2]. Network traffic data contains features that are qualitative or quantitative nature and has to be treated differently [3]. This work studies the effect of normalization on Naive Bayes and J48 Decision tree classifier with the corrected KDDCUP99 and Kyoto 2006+ dataset. A comprehensive approach for normalization for network traffic attributes has been proposed**.**

Keywords-*Intrusion, anomaly, network traffic, normalization, KDDCUP 99, KYOTO 2006+, Naïve Bayes, J48 Decision tree, classifier, pre-process*

*****

I. Introduction and Backgroud

There is high risk of Intrusions because of the increased use of Internet. Network security has become more important as Internet has become the part and parcel of everyone'd day to day activities. Protecting data/information from hackers/intruders is a new area in Computer science [6]. Systems deployed in Internet needs to be protected against various attacks. Organizations have incurred huge losses due to intrusions. Intrusion is a deliberate, unauthorized, illegal attempt to access, manipulate or taking possession of Information System to render them unreliable or unusable. Intrusion Detection is the process of identifying various events occurring in a system/network and analyzing them for possible presence of Intrusion [3]. Various techniques such as soft computing, data mining, statistical methods, machine learning, bio-inspired and artificial intelligence etc., have been used for intrusion detection. For the above methods to work, Data normalization is a fundamental preprocessing step for mining and learning from data [4]. The objective of this work is to study the effect of normalization on the performance of Naive Bayes intrusion detection classifier.

The mean-range normalization executes a linear transformation on original data values and SoumyaParihar et al have normalized the data using mean-range normalization. The authors of [13] claim that the Detection system proposed by them achieves maximum detection accuracy with non-normalized data with few exceptions of attacks and the authors claim that this can be solved using statistical normalization by eliminating bias in the data. Mehar Salem et al have proposed hybrid normalization where the discrete variables are converted using probability mass function and the numerical attributes are normalized using the known normalization methods such as decimal normalization, mean-range normalization or statistical normalization and then these attributes are combined before fed into the classifier [14].

D. Davidson et al. have proposed normalization in the context of protocol type i.e. normalization for context free and context sensitive grammars [15] and claims that normalizers incur an acceptable level of overhead approximately 15% in worst case. Wei Wang et al claim that their experiment shows that the attribute normalization improves the detection performance with *k*-NN, PCA and SVM. Statistical normalization yields better performance if the data sample is large and even mean range [0, 1] can also improve the detection performance [16]. Riti Lath et al claims that the data without any pre-processing, yield good results in case of classification and pre-processing of data takes much of execution time [17]. Vaishali R et al claim that the quality of patterns mined and the time for mining can be substantially improved if the data is pre-processed before mining and have used min-max, z-score and decimal scaling data normalization techniques [18]. The results presented by Chandrasekhar et al. [19] prove that Z-score normalization good terms of detection accuracy and Mean-range normalization is good in terms of execution time.

The organization of the paper is as follows Section 2 discusses various normalization techniques. In section 3 the dataset used in this study, data pre-processing, and the experiments are discussed. The experiment and the results are discussed in section 4. Conclusions and future work in given in section 5.

II. NORMALIZATION TECHNIQUES

Data pre-processing is the first step in analyzing any data. As discussed earlier, Network data consists of both numerical and categorical values. The numerical values and the categorical values need to be treated differently. This section gives the brief explanation of the normalization techniques used in this study. In this paper we use four schemes for quantitative attribute normalization and the two schemes for qualitative attribute normalization.

Quantitative data/attributes can be directly normalized whereas the qualitative data needs to be converted to numeric values before applying any normalization. The dataset which is considered has two 2 qualitative attributes i.e. flag and service and all the other 12 attributes are quantitative.

For Qualitative data the general approach is to replace the values with sequence numerical values such as 1, 2, etc. Though this seems simpler, it does not consider the semantics of the qualitative attributes and we refer this as sequence normalization of qualitative values.

As an alternate way the following probability function is used to normalize the qualitative data.

$$\boldsymbol{fx}\ (\boldsymbol{x}) = \boldsymbol{Pr}\ (\boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{Pr}(\{\ \boldsymbol{s} \in \boldsymbol{S}: \boldsymbol{X}(\boldsymbol{s}) = \boldsymbol{x}\})\ [8] \rightarrow \mathbf{(1)}$$

Based on the above equation the qualitative values are transformed into quantitative values within the range of [0-1]. The qualitative attributes are normalized using the above two schemes. For quantitative date the following four schemes are used.

A. Frequency Normalization

Frequency normalization normalizes an attribute by dividing it with the summed value of the attribute. It is defined as

$$xi = \frac{vi}{\sum_1^n \quad vi} \qquad \rightarrow (2)$$

Frequency normalization scales an attribute between [0, 1].

B. Maximize Normalization

Maximize normalization normalizes an attribute by dividing it with the maximum value of the given attribute. It is defined as:

$$xi = \frac{vi}{\max(vi)} \qquad \rightarrow (3)$$

C. Mean Range Normalization

If the maximum and minimum value of a given attribute is known, it is easy to transform the attribute into a range of value [0, 1] by

$$\boldsymbol{xi} = \frac{\boldsymbol{vi - min(vi)}}{\boldsymbol{max(vi) - min(vi)}} \qquad \rightarrow (4)$$

D. Rational Normalization

Rational normalization is based on the rational function. For each value of an attribute, 1 is divided by the attribute value. It is defined as

$$xi = \frac{1}{vi} \quad \rightarrow (5)$$

III. DATASETS AND EXPERIMENTAL SETUP SETUP

Kyoto 2006+ dataset and Corrected KDD CUP 99 datasets are used to study the effect of normalization on intrusion detection using Naive Bayes and J48 classifier.

A. KYOTO2006+ DATASET

Kyoto 2006+ [8] dataset is a Network Intrusion Evaluation/Detection dataset which was obtained from various honeypots from November 2006 to August 2009. Real network traffic traces were captured in this dataset. This data has 24 statistical features which includes 14 conventional features which were there in KDDCUP '99 Dataset and 10 additional features for effective investigation. This study uses 31st Aug 2009 data and has used the first 14 features (conventional features) and the label which indicates whether the session is an attack or not. As this study does not distinguish between the known and unknown attack, both are represented as attack only. The unknown attacks in this dataset are very minimal and that is also another reason for not distinguishing known and unknown attack. The list of features is given below.

B. KDDCUP 99 Dataset

MIT Lincoln Lab developed a public repository dataset named DARPA KDD Cup '99 (KDD99) Intrusion Detection dataset to promote research in the area of Intrusion Detection. This KDD99 dataset is being used by various researchers. This dataset is based on 1998 DARPA [9] initiative to provide designers of IDS with a benchmark on which various methodologies can be tested. The same was used in International Knowledge Discovery and Data Mining Tools Contest which was held in conjunction with KDD99, the Fifth International Conference on Knowledge Discovery and Data Mining to evaluate the performance of various Intrusion Detection methods [7]. KDD99 Intrusion Detection dataset consists of three subsets namely "10% KDD", "Corrected KDD" and "Whole KDD". The main reason for using the dataset in majority of our evaluations was the need of relevant data that can easily be shared with other researchers, allowing them to duplicate and improve our results. The KDD99 has 41 features or attributes that have either continuous or discrete values and are divided into three groups namely basic features, content features and statistical features which is either time bound or host based features. In this research we have chosen "Corrected KDD" dataset which has 37 types of attacks. In this study only 14features (given below) and the label which indicates whether the session is an attack or not were only used and the other features were not used. The features which were used are given below. The reason for selection these features are that the same features are used in Kyoto 2006+ dataset also.

List of features/attributes used in this study for both KDDCUP 99 and Kyoto 2006+ dataset

- *duration: length (number of seconds) of the connection*
- *service: network service on the destination, e.g., http, telnet, etc.*
- *src_bytes: number of data bytes from source to destination*
- *dst_bytes: number of data bytes from destination to source*
- *count: number of connections to the same host as the current connection in the past two seconds*
- *same_srv_rate: % of connections in the count feature to the same service*
- *serror_rate: % of connections in the count feature that have ``SYN'' errors*
- *srv_serror_rate: % of connections whose service type is the same to that of the current connection in the past two seconds that have "SYN" errors*
- *dst_host_count: among the past 100 connections whose destination IP address is the same to that of the current connection, the number of connections whose source IP address is also the same to that of the current connection*
- *dst_host_srv_count: the number of connections in the dst_host_count feature whose service type is also the same to that of the current connection*
- *dst_host_same_src_port_rate: % of connections in the dst_host_count feature whose source port is the same to that of the current connection*
- *dst_host_serror_rate: % of connections in the dst_host_count feature that have "SYN"*
- *dst_host_srv_serror_rate: % of connections in the dst_host_srv_count feature that "SYN" errors*
- *flag: normal or error status of the connection*
- *label: indicates whether the session is an attack or not*

C. Experimental Setup

The experiments were carried out on a system with Intel Core i3 CPU M 380 @ 2.53 Ghz and 4GB RAM running Window 8 Professional 64-bit Operating System. Microsoft Office Professional Plus 2010 was used for data pre-processing. In this study two schemes for qualitative data and four schemes for quantitative data are proposed. The qualitative attributes 'flag' and 'services are replaced with sequence numbers and the quantitative attributes are normalized using the four schemes as described in earlier section. Similarly, qualitative attributes 'flag' and 'services are replaced with probability function given in Equation 1 and the quantitative attributes are normalized using the four schemes thus, resulting in eight sets of data for each dataset. The study uses two different datasets (KYOTO 2006+ and KDD CUP 99) thus sixteen experiments have to be carried out for this study.

This research uses the basic 'Naive Bayes' classifier and J48 decision tree classifier with the 10-fold cross validation option to test the effect of normalization on Naive Bayes and J48 classifier.

The Naive Bayes classifier is heavily simplified Bayesian probability model [5, 10]. Unlike other classifiers, a single scan of the training data is enough and missing attribute values can be easily handled [11]. Naïve Bayes is a simple classification scheme, in which the class-conditional probability is estimated with the assumption that the attributes are conditionally independent [11]. In other words, Naïve Bayes is a supervised learning algorithm based on Bayes' Theorem with the 'Naïve' assumption that the features are strongly independent and mathematically this is given in Equation 6.

$$P(X1, \dots, Xn|Y) = \pi\ P(Xi|Y) \qquad (6)[10]$$

J48 classifier is a simple C4.5 decision tree for classification. With this technique, a binary tree is constructed to model the classification process. The data is divided into range based on the values of attributes. Once the tree is built, it is applied to each tuple in the database and results in classification for that tuple [12].

In 10-fold cross-validation, the available data is randomly divided into 10 disjoint subsets of approximately equal size and one of the subset is used as the test set and the remaining 9 sets are used for building classifier. The test set is used to estimate the accuracy and this is done repeatedly 10 times so that each subset is used as test set once. Cross validation has been tested extensively and has been found to work well when the sufficient data is available.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

Experiments were carried out for both the datasets described above and the percentages of correctly classified instances are given below Table 1.

TABLE I. PERCENTAGE OF CORRECTLY DETECTED INSTACNES IN CORRECTED KDDCUP99 DATASET

Normalization Technique for Quantitative Variables	Classifier	Sequence Normalization for Qualitative Variables	Probability function Normalization for Qualitative Variables
Frequency	Naïve Bayes	92.2811	92.2811
	J48	95.2419	95.2419
Maximize	Naïve Bayes	92.2811	92.2811
	J48	92.7256	97.7256
Mean Range	Naïve Bayes	91.9065	92.2811
	J48	97.6806	97.7256
Rational	Naïve Bayes	92.2811	93.6569
	J48	97.7256	98.0664

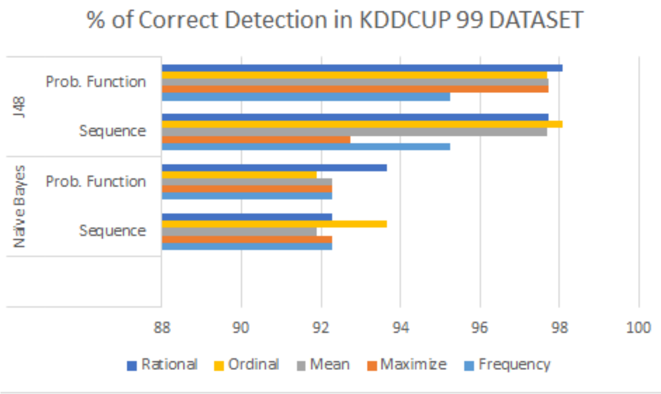

Figure 1. Percentage of correctly detected instances in Corrected KDDCUP 99 dataset.

From above Table 1 and Figure 1, it can be observed that Probability function Normalization gives better results than Sequence Normalization. Probability Normalization for qualitative attributes and Rational Normalization for quantitative attribute produces better results for Corrected KDDCUPP 99 dataset with respect to correctly detected instances

TABLE II. PERCENTAGE OF CORRECTLY DETECTED INSTACNES IN KYOTO 2006+ DATASET

Normalization Technique for Quantitative Variables	Classifier	Sequence Normalization for Qualitative Variables	Probability function Normalization for Qualitative Variables
Frequency	Naïve Bayes	81.0931	81.4117
	J48	95.9425	95.9403
Maximize	Naïve Bayes	81.0916	81.4117
	J48	97.0126	97.0208
Mean Range	Naïve Bayes	79.4119	81.4117
	J48	96.9532	97.0208
Rational	Naïve Bayes	81.0916	72.1962
	J48	97.0126	97.3312

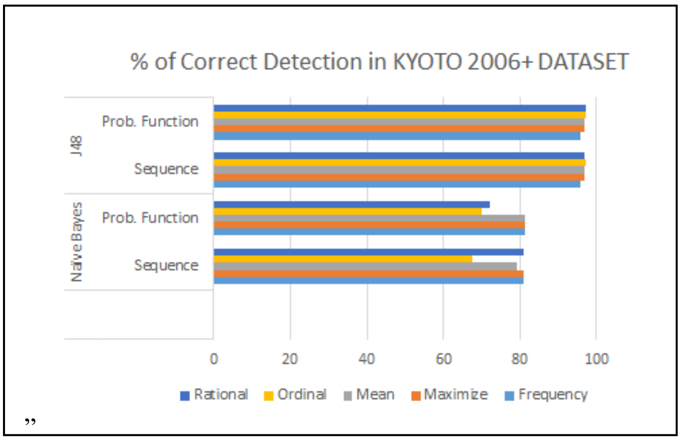

Figure 2. Percentage of correctly detected instances in KYOTO 2006+ dataset.

From the above Table 2 and Figure 2, it can be observed that Probability function Normalization produces better results for all the test cases for Naïve Bayes (except for Frequency Normalization), where the percentage of correct detection is marginally low and J48 (except for Rational Normalization). With regard to KYOTO 2006+ dataset Mean-Range normalization produces good results.

However, the above results pertains only to the percentage of correctly detected instances and the other performance measures such as Detection Rate, Accuracy, False Alarm Rate and F-Score needs to be considered. With regard to the time, there shall not be any differences as the number of tuples and the number of columns in the dataset remains same and only the values are normalized to avoid dominance of extreme values

V. CONCLUSIONS AND FUTURE WORK

In this study the authors have presented various schemes for normalizing the qualitative and quantitative attributes and the performance of the same with respect to Naïve Bayes and J48 Classifier. Overall results suggests that Mean-Range Normalization for quantitative attributes and Probability Function Normalization for qualitative attributes produces better results in terms of percentage of correctly detected instances. As claimed by D. Davidson et al. the normalization increases the execution time by 15% and it should be investigated whether it worth is to normalize at all given the slight improvements in the detection rate. However the question of normalization of attributes for in-band intrusion detection is still a question mark as the data rate and response time requirements are high. These Normalizations can be applied for out-of-band data and offline data which is used for forensics. Normalization of data can play a vital role in offline classification where the response time or execution time is not critical. The authors will continue this study for other performance measures such as F-Score, Detection Rate and accuracy etc. and the way forward for in-line normalization where the data rate is high.

REFERENCES

[1] Ammar, A., “Comparison of Feature Reduction Techniques for the Binominal Classification of Network Traffic,”Journal of Data Analysis and Information Processing,Vol. *3*(02),2005,pp.11.

[2] Soumya Parihar and Nachiketa Tarasia., “Intrusion Detection using Trigonometric functional link Artificial Neural Networks,”International Journal of Advanced Computational Engineering and Networking, Vol. 1(3), 2013,pp.10-14.

[3] D. Ashok Kumar and Venugopalan, S.R., “A Novel algorithm for Network Anomaly Detection using Adaptive Machine Learning.,”Proc. Advanced Computing and Intelligent Technologies (ICACIE), Dec 2016

[4] Song, J., Takakura, H., Okabe, Y., Eto, M., Inoue, D. and Nakao, K., “A Robust Feature Normalization Scheme and Anomaly-Based IDS,”Proceedings of the 12th International Conference on Database Systems for Advanced Applications,2007.

[5] Panda, M. and Patra, M.R., “Network intrusion detection using naive bayes,”International journal of computer science and network security, Vol. *7*(12), 2007, pp.258-263.

[6] D. Ashok Kumar and Venugopalan, S.R., “Intrusion detection by initial classification-based on protocol type,” International

Journal of Advanced Intelligence Paradigms,Vol. 9(2-3), 2017,pp.122-138.

[7] The UCI KDD Archive: KDD Cup 1999 Data, Information and Computer ScienceUniversity of California, Irvine, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html (1999). Accessed 2 February 2014

[8] Song, J., Takakura, H., Okabe, Y., Eto, M., Inoue, D. and Nakao, K., "Statistical analysis of honeypot data and building of Kyoto 2006+ dataset for NIDS evaluation," Proc. of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for SecurityACM, April 2011,pp. 29-36.

[9] MIT Lincoln Lab., Information Systems Technology Group, The 1998 Intrusion detection off-Line Evaluation Plan. http://www.ll.mit.edu/ideval/files/id98-eval-ll.txt

[10] Neethu, B., "Classification of intrusion detection dataset using machine learning approaches,"International Journal of Electronics and Computer Science Engineering, Vol. 1(3), 2012, pp.1044-51.

[11] Farid, D.M., Rahman, M.Z. and Rahman, C.M., "Adaptive intrusion detection based on boosting and naïve bayesian classifier,"International Journal of Computer Applications, Vol. 24(3),2011.pp.12-19.

[12] Patil, T.R. and Sherekar, S.S., "Performance analysis of Naive Bayes and J48 classification algorithm for data classification," International Journal of Computer Science and Applications,Vol. 6(2),2013, pp.256-261.

[13] Relangi, L.P.K. and Varma, M.K.S., 2015. "Improved Mca Based Dos Attack Detection," IJSEAT, Vo.3(8), pp.293-297.

[14] Salem, M. and Buehler, U., 2012,"Mining Techniques in Network Security to enhance Intrusion Detection Systems,"International Journal of Network Security & Its Applications, Vol. 4(6),2012, pp.51-66.

[15] Davidson, D., Smith, R., Doyle, N. and Jha, S., "September. Protocol normalization using attribute grammars,"InEuropean Symposium on Research in Computer Security,Springer Berlin HeidelbergSep 2009, pp. 216-231.

[16] Wang, W., Zhang, X., Gombault, S. and Knapskog, S.J., "Attribute normalization in network intrusion detection," Proc.International Sypmposium on Pervasive systems, algorithms, and networks (ISPAN), IEEE Press, Dec 2009,pp. 448-453.

[17] Lath R, Shrivastava M, "Analytical Study of Different Classification Technique for KDD Cup Data'99,"International Journal of Applied Information, Systems,Vol. 3(36),2012, pp.5-9.

[18] Patel, V.R. and Mehta, R.G., "Impact of outlier removal and normalization approach in modified k-means clustering algorithm,"IJCSI International Journal of Computer Science Issues,Vol. 8(5),2011.

[19] Chandrashekhar, A.M. and Raghuveer, K., "Improvising an Intrusion Detection Precision of ANN Based Hybrid NIDS by incorporating Various Data Normalization Techniques–A Performance Appraisal,"IJREAT International Journal of Research in Engineering & Advanced Technology,Vol. 2(2),2014, pp.1-7.