Attribute Normalization Techniques and Performance of Intrusion Classifiers: A Comparative Analysis

Zohair Ihsan, Mohd Yazid Idris, Abdul Hanan Abdullah

Faculty of Computing, Universiti Teknologi Malaysia, 81310, Skudai, Johor, Malaysia
izohir2@live.utm.my, yazid@cs.utm.my, hanan@utm.my

Abstract: Network traffic have several attributes with different range of values. These attributes can be qualitative or quantitative in nature. Attributes with large values significantly influence the performance of intrusion classifier making it bias towards them. Attribute normalization eliminates such dominance of the attributes by scaling the values of all the attributes within a specific range. The paper discusses various normalization techniques and their influence on intrusion classifiers such as Random Forest, Bayes Net, Naive Bayes, NB Tree and Decision Tree. Furthermore, the concept of hybrid normalization is applied by normalizing the qualitative and quantitative attributes differently. Experiments on KDD Cup 99 suggests that the hybrid normalization can achieve better results as compared to conventional normalization.

[Ihsan Z, Idris MY, Abdullah AH. **Attribute Normalization Techniques and Performance of Intrusion Classifiers: A Comparative Analysis**. *Life Sci J* 2013;10(4): 2568-2576] (ISSN:1097-8135). http://www.lifesciencesite.com. 343

Keywords: Intrusion detection, attribute normalization

1. Introduction

Computer security plays an important role in our society by securing and protecting unauthorized access to data. With ever growing use of computer technologies, wider avenues of research and innovation became a possibility. Despite of these positive aspects the excessive use of the technology made humans more dependent on technology than ever before. Generally the excessive use and dependency on technology made it vulnerable to various risks (Martin 1996). By way of illustration several risk situations could arise if computer is left unattended. For instance anyone can access it and install rootkit or some malicious software. The situation could be much worse if computer is connected to a network, or Internet. As a result of such vulnerability a need stems to secure the computer systems by a security mechanism that will responds to any attacks. One of the notable security mechanism that is currently in use is Intrusion Detection.

Intrusion Detection is a field of computer security that focuses on the detection of intrusive attempts in a computer system. It can be defined as "The process of identifying that an intrusion has been attempted, is occurring, or has occurred" (NSTAC, 1997). Adding further to the knowledge of intrusion detection (Denning, 1987) mentioned that intrusive activities could be identified as they follow different patterns from normal ones. In general, Intrusion Detection Systems (IDS) are classified in two categories: signature-based intrusion detection and anomaly-based intrusion detection. Signature-based intrusion detection tries to looks for attacks by matching it with the known attack signature database whereas Anomaly-based intrusion detection typically relies on knowledge of normal behavior and identifies any deviation from it.

A contributor to the success of IDS is a set of meaningful attributes that are extracted from network traffic. These attributes represents qualitative and quantitative aspects of network traffic. Attributes representing qualitative aspects are nominal in nature, while attributes representing quantitative attributes are numerical in nature. For instance, in KDD Cup 99 (The KDD Archive, 1999), "service" attribute has nominal values whereas, "duration" attribute has numerical values. Qualitative attribute can be transformed in to quantitative attribute by assigning numerical values. Such transformation are performed when the intrusion classifier can only process qualitative data. On the other hand, quantitative attribute have numerical values which are based on well-defined scale in discrete or continuous order. In KDD Cup 99, "serror_rate" ranges from [0-1] while "dst_bytes" ranges from [0-5155468]. The variance in range of quantitative attributes significantly affect the performance of intrusion detection classifier by making it bias towards the attribute with large values. To eliminate such bias from intrusion classifier, the values of the quantitative attributes are normalize by scaling them with in a specific range.

The objectives of this work is to study the performance of an intrusion classifier over different attribute normalization techniques. Attribute normalization is a first and critical step in the domain of intrusion detection. To measure the effect of attribute normalization techniques, four normalization techniques namely Frequency

Normalization, Maximize Normalization, Mean Range Normalization and Rational Normalization are used to normalize KDD Cup 99 data. These normalization techniques scale the data within the range of [0-1]. The normalized data is evaluated by classifiers like Random Forest (Breiman, 2001), Bayes Net (Friedman et al., 1997), Naive Bayes (Pearl, 1988), NB Tree (Kohavi, 1996) and Decision Tress (Machine Learning, 1997) for the classification of intrusions. Furthermore, the concept of hybrid normalization is applied by adopting separate attribute normalization for qualitative and quantitative attributes. The hybrid normalization is a two-step normalization process. In the first step quantitative attributes are normalized using Mean Range Normalization, Frequency Normalization, Maximize Normalization and Rational Normalization In the second step, qualitative attributes are normalized using probability.

The remainder of paper is organized as follows. Section 2 covers the literature review. Section 3 explain the normalization techniques for qualitative and quantitative attribute. Experimental details are given in section 4. Section 5 covers the discussion and concluding remarks follows in Section 6.

2. Literature Review

The work of (Chakraborty and Chakraborty, 2000) used dimension space to normalized only quantitative attributes. Attributes are transform to higher dimension space till all the attributes are at same level in space. They normalized using the longest feature vector and divided all the attributes by it. (Cai et al.,2010) proposed unified normalization distance framework for qualitative and quantitative attributes using distance function between the attributes, they mapped the qualitative values to a categories domain so that coordination of qualitative value is 1 in its dimension in the real number space. (Yu Liping et al.,2009) work compared various normalization techniques in multi attribute environment and conclude that different evaluation methodology requires different data normalization.

The work of (Oh et al., 2009 include the transformation of the “protocol type” and “service” attribute with corresponding decimal number as defined in IANAs “Assigned Internet Protocol Numbers” (IANA,2013). They used mean range to normalized the data. Adopting IANA assigned protocol numbers for transforming the nominal values with numeric values can affect the normalization and classification results since these number were defined in a different context. There are many services on Internet, each has its importance in a given scenario. For instance in an organization email (SMTP Port 25, POP3 Port 110) is more important than web browsing (HTTP Port80), whereas for an average internet user the case may be vice versa. Transforming the nominal values to its identification number without any mapping function can result in false classification.

Work by (Ippoliti and Zhou ,2010) presented a Growing Hierarchical Self Organizing Map (GHSOM) normalization method using mean range normalization approach. They proposed a dynamic process for nominal attributes. (Wang et al., 2009) presented comparison four normalization techniques which scaled the data within different ranges. Comparison of normalization techniques scaling in different range of normalize data can give inaccurate results. The work of (Said et al.,2011) comparatively studied the normalization methods in unsupervised learning using Principle Component Analysis (PCA). Their results shows that log normalization is the best for PCA, however they don’t address the qualitative attributes in their work. In (Brifcani, A. and A. Issa , 2011), comparative study of supervised learning using decision tree on nominal attributes have been investigated. (Hernndez et al.,2006) suggested three different methods for transforming qualitative attributes using support vector machine(SVM) and feed forward neural network.

3. Attribute Normalization

Data pre-processing is the first step whilst analyzing the data. The data pre-processing is comprised of phases like dataset creation, data cleaning, integration, feature construction, normalization, feature selection and discretization (Kotsiantis et al., 2006). (Kurgan and Musilek, 2006) reported that data pre-processing can take up to 50% of the overall process effort. This section provides a brief explanation of the normalization phase of data pre-processing.

A dataset is collection of different attributes that describes various characteristics of the data. These attributes can be qualitative or quantitative in nature with different range of values. The nature and values of these attributes influences the data analysis process. For instance, Attributes with large values can dominate attributes with small value. The process of normalization can eliminate such dominance by scaling them all within a specific range. Quantitative attributes can be directly normalized, whereas in case of qualitative attributes, the nominal values first needs to be converted to numeric value before applying the normalization. The numeric values can be assigned based on certain criteria or simply replacing every nominal value with 1,2,3…n. Once the qualitative attributes have been converted to

quantitative attributes, the normalization process can be applied to them. The normalization techniques discussed in section 3.1 to section 3.4 requires the qualitative data to be transformed in to quantitative, whereas the hybrid normalization technique section 3.5 discusses do not requires such transformation of qualitative attributes.

3.1 Mean Range Normalization:

The mean range normalization normalizes an attribute value by subtracting minimum value of that attribute from the current value. This value is further divided by the difference maximum and minimum value of that attribute. It is define as:

$$x_i = \frac{v_i - \min(v_i)}{\max(v_i) - \min(v_i)}$$

3.2 Frequency Normalization:

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

$$x_i = \frac{v_i}{\sum v_i}$$

3.3 Maximize Normalization:

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

$$x_i = \frac{v_i}{\max(v_i)}$$

3.4 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:

$$x_i = \frac{1}{v_i}$$

3.5 Hybrid Normalization:

As discussed earlier, attribute normalization scales the value of all the attributes within a specific range. For quantitative attributes, we can directly applied the normalization function and scale them. However, qualitative needs to be transform to quantitative values before the normalization can be applied to them. A general approach is to replace the qualitative values with quantitative/numeric values. This approach is simple but it neglects the semantics of these qualitative attributes. In hybrid normalization, the probability function is used to normalize the qualitative attributes. Suppose that X is a qualitative attribute define as a X= a,d,b,a,c,d,b,a,b,d,d,c , where N=12 and K=4. Using probability function $f_x(x)$.

$$f_x(x) = \Pr(X = x) = \Pr(\{s \in S : X(s) = x\})$$

Based on the function $f_x(x)$, the qualitative values are transformed in to quantitative values within the range of [0-1] $f_x(a)$=3/12=0.25, $f_x(b)$=3/12=0.25, $f_x(c)$=2/12=0.16, $f_x(d)$=4/12=0.33. In case of real time intrusion detection, the sliding windows can be adopted for normalizing of qualitative attributes. Figure. 1 shows flow charts the conventional and hybrid normalization approaches.

4. Experiments

Experiments are performed using KDD Cup 99 data sets. (Tsai et. al., 2009) mentioned that this dataset has been used in 30 major IDS studies. This is the largest, publically available, dataset for researchers in the field of network intrusion detection (Raghuveer, 2012; Wu and Banzhaf, 2010; Zhao et al., 2013). The KDD Cup 99 was generated from the 1998 DARPA Intrusion Detection Evaluation Program (IDEP) (Lippmann et al., 2000) prepared by MIT Lincoln Lab DARPA IDEP contains 9 weeks of raw tcpdump data collected from simulation of network attacks having 7 weeks of training data (5 million records) and 2 weeks of testing data (2 million records). KDD Cup 99 dataset has total 41 attributes (9 connection based, 13 content based and 19 time based). Out of these 41 attributes, 34 are quantitative while 7 are qualitative in nature.

Experiments are performed on subset of 10% KDD Cup 99 having 12678 records. Table 1 shows the record types in in actual 10% KDD Cup 99 and the subset created for the experiments. On the other hand the Table 2 comprises the list of attacks in 10% KDD Cup 99 dataset and subset respectively. The subset comprises of 5000 Normal connection, 4000 DoS attacks (Denial of Service), 2500 Prob attacks (Surveillance and Monitoring), 1126 R2L attacks (Unauthorized access to a remote machine) and 52 U2R attacks (Unauthorized accessing higher privilege user account).

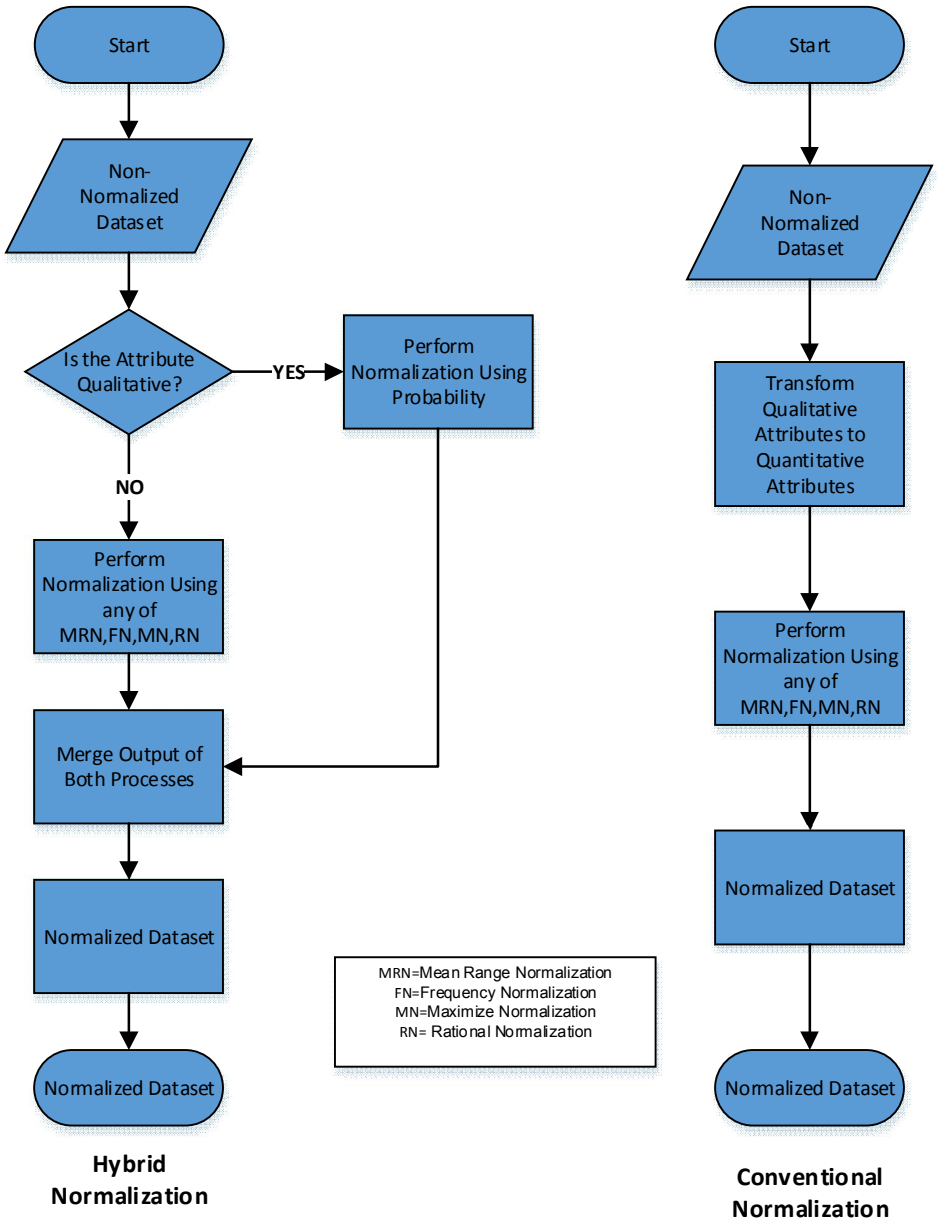

Figure 1. Attribute Normalization Approaches

Table 1. KDD CUP 99 Datasets

Record Type	10% KDD	Subset Used in experiments
Normal	97277	5000
DoS	391458	4000
Prob	4107	2500
R2L	1126	1126
U2R	52	52
Total	494020	12678

Table 2. List of Attacks in KDD CUP 99

Attack Type	10% KDD	Subset Used in experiments
back.	2203	183
buffer_overflow.	30	30
ftp_write.	8	8
guess_passwd.	53	53
imap.	12	12
ipsweep.	1247	784
land.	21	1
loadmodule.	9	9
multihop.	7	7
neptune.	107201	2
nmap.	231	231
normal.	97278	5000
perl.	3	3
phf.	4	4
pod.	264	20
portsweep.	1040	278
rootkit.	10	10
satan.	1589	1207
smurf.	280790	3695
spy.	2	2
teardrop.	979	99
warezclient.	1020	1020
warezmaster.	20	20
Total	494021	12678

Table 3 shows the qualitative attributes of KDD Cup 99.These include the "protocol_type" (type of protocol used). "Service" (name of network service on the destination machine). "Flag" (normal or error status of flag values). The other 4 qualitative attributes are binary in nature having value of 1 and 0. The value of "Land" is 1 if the connection is from/to the same host port else it will be 0. The "logged-in" will be 1 If the login is successful and vice versa. The value of "is_host_login" and "is_guest_login" will be 1 if the login is a host or guest respectively else it will be 0. Since the first 3 qualitative attributes are enriched with respect to data dimensionality, the hybrid normalization is applied only to them whereas binary conversion is applied to other 4 attributes.

Experiments setup involves creating 8 identical datasets each consisting of 12,678 records from 10% KDD Cup99. In the first set of experiments, the qualitative attributes in the first 4 dataset were transformed by assigning numeric values in continuous manner. In the next step, each dataset is normalized using the normalization technique namely Mean Range Normalization, Frequency Normalization, Maximize Normalization and Rational Normalization. Each normalized data set is evaluated using Random Forest, Bayes Net, Naive Bayes, NB Tree and Decision Tree respectively and their performance is measured in term of Intrusion Classification Rate (ICR) and Model Building Time (MBT) using 10 fold cross validation.

Table 3. KDD CUP 99 Qualitative Attribute

Attribute ID	Attribute Name	Number of Values
f2	protocol_type	3
f3	service	67
f4	flag	11
f7	land	2
f12	logged-in	2
f21	is_host_login	2
f22	is_guest_login	2

In the second set of experiments, last 4 dataset were normalized using hybrid normalization. Qualitative attributes were normalized using probability function in all 4 datasets, while quantitative attributes were normalized using the normalization technique namely Mean Range Normalization, Frequency Normalization, Maximize Normalization and Rational Normalization. Same performance measurement parameters were used. Table 4 and Table 5 shows the results of ICR and MBT

5. Discussion

Results of the study shows that data normalization process have significantly influenced the performance of intrusion classifier not only in terms of intrusion classification rate but also in terms of processing time. Among the intrusion classifier, Random Forest achieved the highest ICR for all normalization techniques, both in case of conventional normalization and hybrid normalization. Whereas, the ICR of Naïve Bayes is the lowest. Hybrid normalization has shown improvement of ICR in 15 experiments out of total 20 experiments and overall ICR increased by 2.45% as compared to conventional normalization. However there was a significant improvement in results of Decision Tree using Frequency Normalization. The ICR of Decision Tree for conventional normalization was 63.30% which increased to 99.27% for hybrid normalization. Similarly NB Tree and Naïve Bayes have shown improvement in hybrid normalization for Frequency Normalization and Rational Normalization. When considering the normalization techniques, the average ICR for Mean Range Normalization is highest (97.50%) in conventional normalization. Whereas in case of hybrid normalization, Rational Normalization has highest (97.85%) average ICR Table 4 and Figure. 2 show details of Intrusion Classification Rate. If we look at the model building time, NB Tree is the most time intensive classifier with good ICR.

The MBT for Naïve Bayes is lowest in every normalization technique for conventional and hybrid normalization, however when we look at the ICR of Naïve Bayes, it average 90%. Although the overall MBT for hybrid normalization is 17.74 seconds more than conventional normalization due the MBT of NB Tree. If we exclude the NB Tree, then there is an improvement of 5.38 seconds for hybrid normalization. Table 5, Figure. 3 and Figure. 4 show the details model building time.

Table 4. 10 Fold Cross Validation of Dataset Using Conventional and Hybrid Normalization Technique

Normalization Technique	Classifier	10 Fold Cross Validation of Classification	
		Conventional Normalization	Hybrid Normalization
Frequency Normalization	Random Forest	96.96%	99.68%
	Bayes Net	95.52%	98.34%
	Naive Bayes	90.38%	90.81%
	NB Tree	96.69%	99.60%
	Decision Tree	63.30%	99.27%
	Average ICR of FN	88.57%	97.54%
Maximize Normalization	Random Forest	99.69%	99.74%
	Bayes Net	98.58%	98.56%
	Naive Bayes	90.23%	90.51%
	NB Tree	99.56%	99.60%
	Decision Tree	99.35%	99.62%
	Average ICR of MN	97.48%	97.61%
Mean Range Normalization	Random Forest	99.76%	99.74%
	Bayes Net	98.58%	98.56%
	Naive Bayes	90.24%	90.51%
	NB Tree	99.55%	99.60%
	Decision Tree	99.35%	99.62%
	Average ICR of MRN	97.50%	97.61%
Rational Normalization	Random Forest	99.72%	99.68%
	Bayes Net	98.80%	98.82%
	Naive Bayes	88.65%	91.69%
	NB Tree	99.51%	99.57%
	Decision Tree	99.60%	99.55%
	Average ICR of RN	97.26%	97.86%

Table 5. Model Building Time of Dataset Using Conventional and Hybrid Normalization Technique

Normalization Technique	Classifier	Time in Seconds	
		Conventional Normalization	Hybrid Normalization
Frequency Normalization	Random Forest	6.55	3.70
	Bayes Net	0.81	0.80
	Naive Bayes	0.46	0.39
	NB Tree	35.69	67.33
	Decision Tree	2.50	1.35
Maximize Normalization	Random Forest	2.26	1.98
	Bayes Net	1.24	1.00
	Naive Bayes	0.47	0.37
	NB Tree	188.83	173.06
	Decision Tree	2.12	1.81
Mean Range Normalization	Random Forest	2.07	1.93
	Bayes Net	0.85	0.75
	Naive Bayes	0.41	0.37
	NB Tree	159.97	181.51
	Decision Tree	1.74	1.60
Rational Normalization	Random Forest	1.92	1.90
	Bayes Net	0.69	0.92
	Naive Bayes	0.55	0.44
	NB Tree	125.00	110.71
	Decision Tree	1.84	1.79

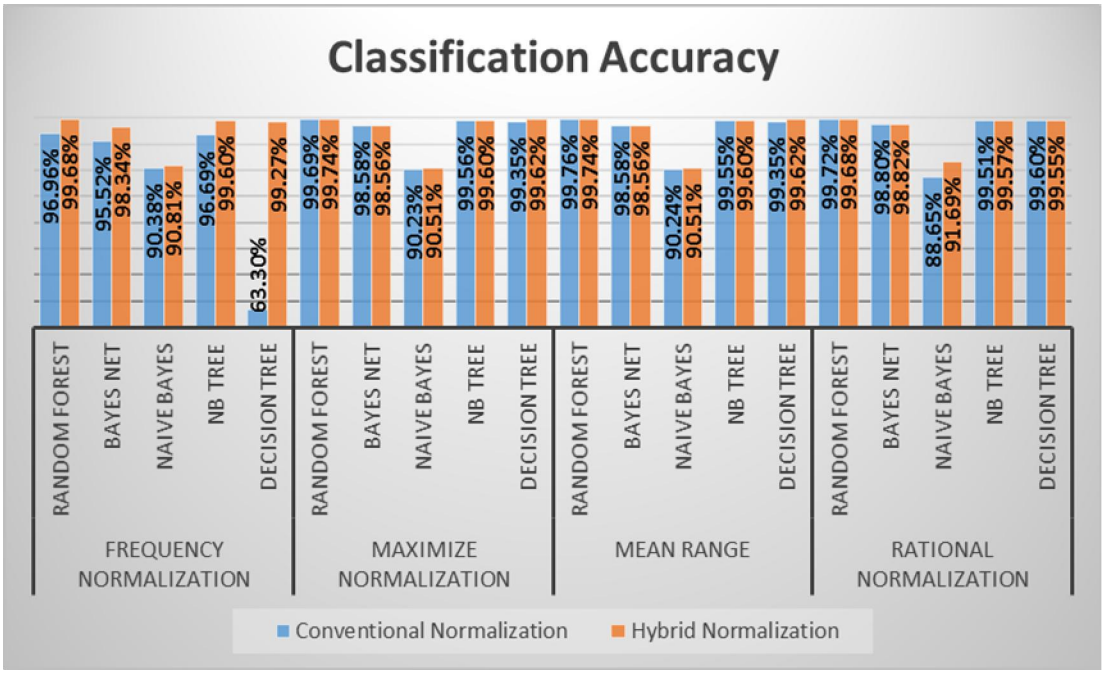

Figure. 2. Intrusion Classification of Conventional Normalization and Hybrid Normalization

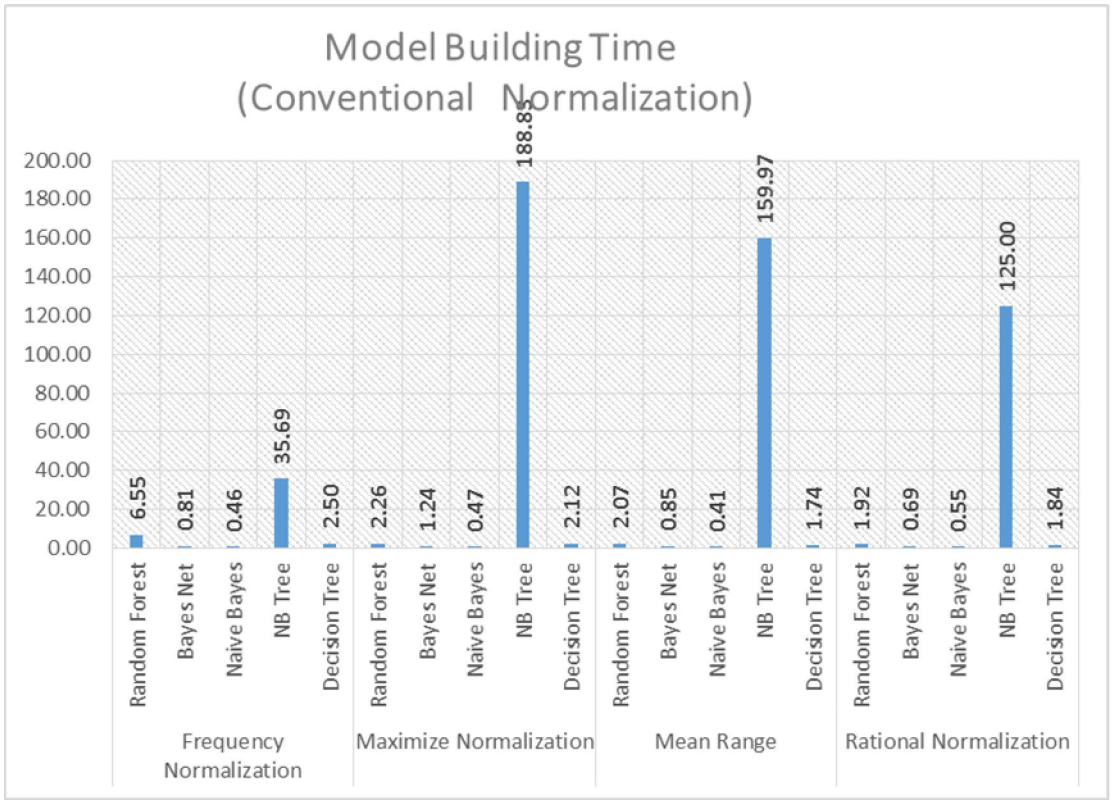

Figure. 3. Model Building Time for Conventional Normalization

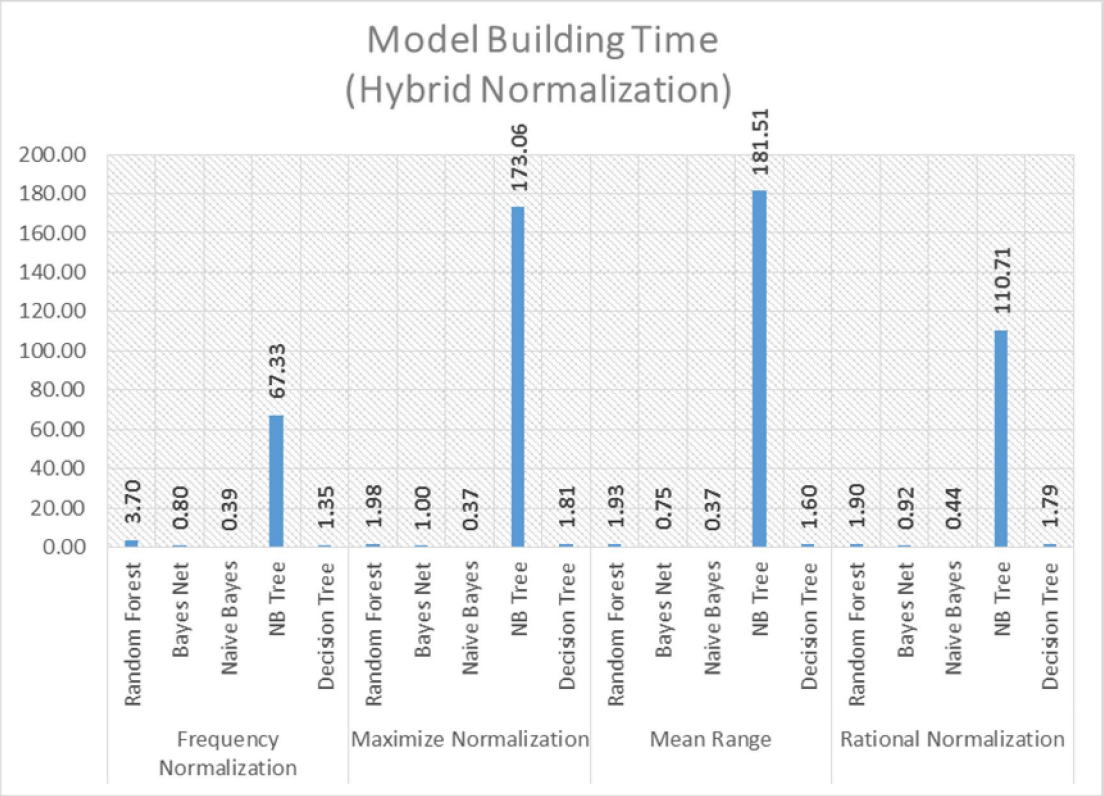

Figure. 4. Model Building Time for Hybrid Normalization

6. Conclusion

Intrusion classification and detection is an active research area in the field of network security. Pre-processing of data for intrusion classifier is a critical step which can significantly affect the performance of intrusion classifier. This paper is an attempt to measure the performance of different intrusion classifier using different normalization techniques. Results of experiment suggests that a higher intrusion classification rate can be achieve with minimum processing time using Mean Rang Normalization technique with Random Forest and applying Hybrid Normalization.

Acknowledgements:

This research is supported by Research Management Center (RMC) of Universiti Teknologi Malaysia.

Corresponding Author:
Mohd Yazid Idris and Abdul Hanan Abdullah
Faculty of Computing
Universiti Teknologi Malaysia
81310, Skudai, Johor, Malaysia
E-mail: yazid@cs.utm.my, hanan@utm.my

References

1. B. Martin, "Technological vulnerability" Technology in Society, vol. 18, no. 4, pp. 511 – 523, 1996.
2. NSTAC, Network Group Intrusion Detection Subgroup Report: "Report on the NS/EP Implications of Intrusion Detection Technology Research and Development", December 1997.
3. D. E. Denning, "An Intrusion-Detection Model" IEEE Transactions on Software Engineering, vol. 13, no. 2, pp. 222–232, February 1987.
4. Stolfo, S. J., "KDD Cup 1999 Dataset." UCI KDD repository, 1999 [Online] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
5. L. Breiman, "Random Forests," Machine learning, vol. 45, no. 1, pp. 5–32, 2001.
6. N. Friedman, D. Geiger and M. Goldszmidt, "Bayesian network classifiers" Machine learning, vol. 29, no. 2-3, pp. 131–163, 1997.
7. J. Pearl, "Probabilistic reasoning in intelligent systems: networks of plausible inference", Morgan Kaufmann Publishers Inc., 1988.
8. R. Kohavi, "Scaling up the accuracy of naive-Bayes classifiers: A decision tree hybrid" in Proceedings of the second international conference on knowledge discovery and data mining, vol. 7, 1996.

9. T. M. Mitchell, “Machine Learning”, McGraw-Hill, 1997.
10. G. Chakraborty and B. Chakraborty, “A novel normalization technique for unsupervised learning in ANN” Trans. Neur. Netw., vol. 11, no. 1, pp. 253–257, 2000.
11. Cai, L. Z., Chen, J., Ke, Y., Chen, T., and Li, Z. G., “A new data normalization method for unsupervised anomaly intrusion detection”. Journal of Zhejiang University SCIENCE C, vol. 11, no. 10, pp. 778-784, 2010.
12. L. Yu, Y. Pan, and Y. Wu, “Research on Data Normalization Methods in Multi-Attribute Evaluation” in International Conference on Computational Intelligence and Software Engineering, pp. 1-5, 2009.
13. H. Oh, I. Doh, and K. Chae, “Attack Classification Based on Data Mining Technique and Its Application for Reliable Medical Sensor Communication” International Journal of Computer Science and Applications, vol. 6, no. 3, pp. 20–23, 2009
14. “IANA Protocol Numbers.” Updated 2013-11-12 [Online]. http://www.iana.org/assignments/protocol-numbers/protocol-numbers.xml
15. D. Ippoliti and Z. Xiaobo, “An Adaptive Growing Hierarchical Self Organizing Map for Network Intrusion Detection,” in Proceedings of 19th International Conference on Computer Communications and Networks, pp. 1–7, 2010.
16. W. Wei, Z. Xiangliang, S. Gombault, and S. J. Knapskog, “Attribute Normalization in Network Intrusion Detection” in 10th International Symposium on Pervasive Systems, Algorithms, and Networks, pp. 448–453, 2009.
17. D. Said, L. Stirling, P. Federolf, and K. Barker, “Data preprocessing for distance-based unsupervised Intrusion Detection” in Ninth Annual International Conference on Privacy, Security and Trust, pp.181–188, 2011.
18. Brifcani, A. and A. Issa, “Intrusion Detection and Attack Classifer Based on Three Techniques: A Comparative Study” Engineering and Technology Journal, vol. 26, no. 2, pp. 368–412, 2011.
19. E. Hern´andez-Pereira, J. A. Su´arez-Romero, O. Fontenla-Romero, and A. Alonso-Betanzos, “Conversion methods for symbolic features: A comparison applied to an intrusion detection problem” Expert Systems with Applications, vol. 36, no. 7, pp. 10612–10617, 2009.
20. . B. Kotsiantis and D. Kanellopoulos, “Data Preprocessing for Supervised Leaning”, International Journal of Computer Science, vol. 1, no. 2, pp. 111–117, 2006.
21. L. A. Kurgan and P. Musilek, “A survey of Knowledge Discovery and Data Mining process models” Knowl. Eng. Rev., vol. 21, no. 1, pp. 1–24, 2006.
22. Tsai, C. F., Hsu, Y. F., Lin, C. Y., and Lin, W. Y, “Intrusion detection by machine learning: A review”. Expert Systems with Applications, vol. 36, no.10, pp. 11994-12000, 2009.
23. S. X. Wu and W. Banzhaf, “The use of computational intelligence in intrusion detection systems: A review” Applied Soft Computing, vol. 10, no. 1, pp. 1–35, 2010.
24. K. Raghuveer, “Performance evaluation of data clustering techniques using KDD Cup-99 Intrusion detection data set” International Journal of Information and Network Security (IJINS), vol. 1, no. 4, pp. 294–305, 2012.
25. L. Zhao, H. S. Kang, and S. R. Kim, “Improved clustering for intrusion detection by principal component analysis with e□ective noise reduction” in Information and Communication Technology. Springer, pp. 490–495, 2013.
26. R. P. Lippmann, D. J. Fried, I. Graf, J. W. Haines, K. R. Kendall, D. McClung, D. Weber, S. E. Webster, D. Wyschogrod, R. K. Cunningham, and M. A. Zissman, “Evaluating intrusion detection systems: the 1998 DARPA o□-line intrusion detection evaluation” in Proceedings of DARPA Information Survivability Conference and Exposition, vol. 2, pp. 12–26, 2000.

7/1/2013