Analysis of the Effect of Clustering the Training Data in Naive Bayes Classifier for Anomaly Network Intrusion Detection

Uma Subramanian and Hang See Ong

***Abstract*—This paper presents the analysis of the effect of clustering the training data and test data in classification efficiency of Naive Bayes classifier. KDD cup 99 benchmark dataset is used in this research. The training set is clustered using k means clustering algorithm into 5 clusters. Then 8800 samples are taken from the clusters to form the training and test set. The results are compared with that of two Naive Bayes classifiers trained on random sampled data containing 8800 and 17600 instances respectively. The main contribution of this paper is that it is empirically proved that the training set derived from clusters generated by k-means clustering algorithm improves the classification efficiency of the Naive Bayes classifier. The results show the accuracy of the Naive Bayes classifier trained with clustered instances is 94.4% while that of normal instances are 85.41% and 89.26%.**

***Index Terms*—Network security, machine learning, classifier evaluation, anomaly intrusion detection.**

I. Introduction

Network intrusion detection is gaining importance due to the rise in network attacks every year. Network intrusion detection can be classified into two types named signature based and anomaly based network intrusion detection.

Anomaly based Network intrusion detection (ANID) is the only solution for novel attacks on networks. The anomaly based Network intrusion detection methods suffers from large false alarm rate.

The objective of this paper is to analyse the effect of clustering training data in the classification efficiency of the Naive Bayer classifier. KDD 99 benchmark data set is used in all the experiments so that the results reported here can be compared with that of others. The KDD cup 99 data set consists of 4 types of attack data and normal data. Denial of service (Dos), probe, Remote to local (R2L), user to root (U2R) are the 4 type of attack data. KDD 99 data has 41 features. The rest of the paper is organized into related works, Proposed Architecture, Experimental setup and Evaluation criteria, Experimental results and Analysis and Conclusion.

II. Related Works

Research in ANID can be grouped into 3 categories. Research on developing innovative, hybrid or ensemble based classifiers [1]-[4], feature selection techniques [5]-[8], and on the training dataset. Research on dataset is minimal.

C.-F. Tsai and C.-Y. Lin [9] proposed the triangle area based neighbours (TANN) hybrid model. In TANN k-means clustering and K-NN classifiers are used to form the hybrid model. This model gave higher accuracy, detection rate and lower False Alarm Rate (FAR) when compared with the three baseline models based on SVM used in the experiments.

S.-J. Horng *et al.* [2] Proposed novel intrusion detection on hierarchical clustering and support vector machines. The Hierarchical clustering algorithm is used to obtain fewer, abstracted, and higher-qualified training instances from the KDD 99 dataset. This model gave best overall accuracy and better DOS and Probe attack detection rates.

This paper's focus is on reducing the number of instances in the training data set. The reason is if it is possible to reduce the size of training data, the cost of labelling can be reduced. Training a supervised classifier like Naive Bayes requires labelling the training data. According to P. Casas, J. Mazel, and P. Owezarski [10], this task is time consuming and complex. Increase in training data translates into increase in training cost due to labelling for large multidimensional data set like kdd 99 benchmark dataset.

K-means Clustering is employed to select a small and better quality training dataset in this research. This paper attempts to find if comparatively small but better quality instances can increase the classification efficiency of the Naive Bayes classifier.

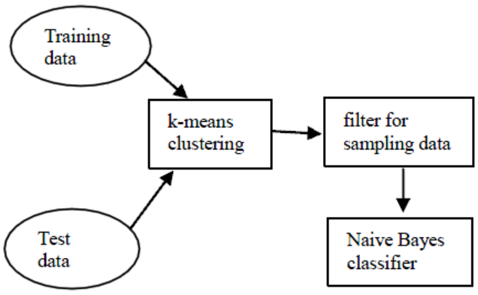

Fig. 1. Architecture of the proposed system.

III. Proposed Architecture

The architecture of the proposed system is shown in Fig. 1. There are two components in the proposed architecture.

Manuscript received July 10, 2013; revised September 22, 2013.

Uma Subramanian and Hang See Ong are with the Department of Electronics and Communication Engineering, College of Engineering, Universiti Tenaga Nasional, Jalan IKRAM- UNITEN, Kajang, Selangor, 43000, Malaysia (Corresponding author: Uma Subramanian, e-mail: umas746@gmail.com).

DOI: 10.7763/JACN.2014.V2.87

One is the unsupervised K-means clustering algorithm and the other is the supervised Naive Bayes Classifier. K-means clustering algorithm is used for clustering the KDD 99 dataset (10%). The clusters are sampled and a new small data set is formed. Naive Bayes classifier is trained and tested using the new data set.

Naive Bayes is a multiclass classifier. Naive Bayes treats all the features independently. All the 41 features of the KDD 99 data set are used in the experiment for classification. H. G. Kayacik, A. N. Zincir-Heywood, and M. I. Heywood [3] have given a detailed description of KDD 99 benchmark dataset. This approach of unsupervised clustering and supervised classification makes the proposed model into a semi supervised learning model. This model will be referred as semi supervised Naive Bayes model (SSNB) in this paper. SSNB has to be compared with a Naive Bayes classifier using normal training set without clustering. This Naive Bayes model will be referred in this paper as Supervised Naive Bayes (SNB). The intrusion detection problem is a 5 class classification problem. The given instance is classified into one of Normal, DOS, Probe, U2R, and R2L classes by the classifier.

IV. Experimental Setup and Evaluation Criteria

A. Experimental Setup

All the experiments are conducted using weka-3-7 software. The experiments were conducted in a machine using Pentium Intel(R) Core(TM) i5 processor and windows 7 Home premium OS. Three datasets namely data set 1, data set 2, and data set 3 is used in the experiments. The three datasets are derived from the original KDD cup 99 benchmark dataset and pre-processed. Data set 1 and data set 2 are used by SNB whereas dataset 3 is used by SSNB. Class distribution of data set 1 containing 8800 randomly selected instances is shown in Table I and that of data set 2 containing 17600 randomly selected instances is shown in Table II. Distribution of Dataset 3 containing 8800 randomly selected instances from the clustered KDD cup 99 dataset is shown in Table III. Cross validation method is used for training and testing both SNB and SSNB. The SNB is trained first using data set 1 containing 8800 instances and the results are tabulated in Table V. Then SNB is then trained with double the amount of instances of SSNB i.e.17600. The results are tabulated in Table VI. The classes normal, probe, DOS, U2R, and R2L are given the labels 0, 1, 2, 3, and 4 respectively in Tables I through III. In Table I, and Table II, the instances are chosen randomly whereas in Table III the KDD 99 dataset is clustered using k-means clustering algorithm into 5 clusters. The clusters are sampled to constitute the data set 3. SSNB is trained and tested using data set 3.The result obtained are listed in Table IV.

TABLE I: CLASS DISTRIBUTION IN DATA SET 1 OF SNB

Class	Number of instances
0	2600
1	2600
2	2600
3	52
4	948

TABLE II: CLASS DISTRIBUTION IN DATA SET 2 OF SNB

Class	Number of instances
0	6240
1	4107
2	6240
3	52
4	961

TABLE III: CLASS DISTRIBUTION IN DATA SET 3 OF SSNB

Class	Number of instances
0	3520
1	1424
2	3792
3	5
4	59

Since the object of this paper is to determine the effect of clustering the training set for the purpose of reducing the number of training instances to reduce labelling cost, the feature selection techniques are not used and all the 41 features of the KDD cup 99 dataset is used.

B. Evaluation Criterion

Classification accuracy alone if not enough to prove that a certain classifier if performing well in comparison to the other. So the other relevant statistical measures must also be considering in doing so. Thus, in this research, Classification accuracy, FPR, DR, and F-measure are used in addition to kappa statistic, relative absolute error and Root Mean Squared Error (RMSE). FAR is the Ratio of misclassified normal connection to total number of connection. ROC curve is not used for evaluation and instead F-measure is used.

DR is the ratio of number of correctly detected attacks to total number of attacks.

F-measure = 2 / ((1/precision) + (1/recall))

As pointed out by M. Tavallaee [11], F-measure uses the properties of both recall and precision. So it is more suitable for evaluation of classifiers. The ideal f-measure value is 1 meaning that there are no false alarms produced. Otherwise it means 0 FPR.

Kappa statistic shows the agreement between predicted and actual classes. Kappa statistic value of 1 indicates the predicted and actual values of the classes are 100% in agreement.

Room mean squared error (RMSE) indicates how precise the classification is. Lower RMSE value indicates the more accurate classifier results [12]. RMSE value is also used in this research for evaluation.

V. Experimental Results And Analysis

The results obtained after classifying the two models SSNB and SNB are given in Table III and Table IV respectively. The percentage of correctly classified instances for the SSNB is 94.3977 % where as that of SNB is 85.4091 %. SSNB has given a high classification percentage when compared to SNB for the same number of training instances.

The F measures of SSNB in Table IV shows that SSNB performs better for all classes except probe when compared with that of SNB as shown in Table V. When using dataset

2, the f-measure of SNB is still better except DOS class.

TABLE IV: CLASSIFICATION RESULTS FOR SSNB

	Normal	Probe	Dos	U2R	R2L
Normal	3301	132	46	28	10
Probe	13	1393	11	7	0
Dos	43	176	3565	7	1
U2R	2	0	0	3	0
R2L	3	7	4	0	45
FPR	0.012	0.043	0.012	0.005	0.001
DR	0.938	0.978	0.94	0.6	0.763
F-measure	0.959	0.889	0.961	0.12	0.783

TABLE V: CLASSIFICATION RESULT FOR SNB USING 8800 INSTANCES

	Normal	Probe	DOS	U2R	R2L
Normal	5696	194	54	268	28
Probe	57	3742	22	286	0
DOS	46	324	5864	5	1
U2R	2	2	0	46	2
R2L	10	58	0	531	362
FPR	0.01	0.057	0.006	0.096	0.002
DR	0.889	0.915	0.935	0.923	0.365
F-measure	0.932	0.893	0.96	0.103	0.529

TABLE VI: CLASSIFICATION RESULT FOR SNB USING 17600 INSTANCES

	Normal	Probe	DOS	U2R	R2L
Normal	2312	144	20	114	10
Probe	19	2379	15	187	0
DOS	15	152	2431	0	2
U2R	1	2	0	48	1
R2L	13	54	0	535	346
FPR	0.01	0.043	0.007	0.062	0.002
DR	0.913	0.911	0.94	0.885	0.377
F-measure	0.945	0.888	0.963	0.077	0.535

The accuracy of SNB when using data set1 and data set 2 are 85.41 % and 89.26% respectively. Even when the number of instances is increased twice to 17600, SNB managed to achieve an accuracy of 89.26% only which is less than 94.4%, the classifier accuracy of SSNB.

The kappa statistic for SSNB is 0.9122. Their corresponding kappa statistics are 0.80 and 0.85 respectively while SSNB recorded 0.91. Fig. 2 shows the kappa statistics value of all the three datasets.

The RMSE values of the SSNB and SNB are plotted in Fig. 3 It shows that SSNB using dataset 3 derived after clustering the KDD 99 data set has the least value. Also it must be noted that the RMSE increases with increase in number of training instances in the normal data sets data set1, and dataset 2.

The normal category or class 0 of SSNB dataset 3 and SNB dataset 2 produced the same FPR of 0.1. But SSNB dataset used 3520 instances whereas SNB dataset 2 used 6240 instances as shown in Tables II and III. Also SSNB dataset has produced higher DR in this category as listed in Table IV and VI.

The probe category or class 1 of SSNB dataset3 and SNB dataset 1 has the same value of 0.043 for FPR. But SSNB produced this value with 3424 probe instances whereas SNB dataset 1 used 2600 instances.

The DOS category or class 2 of SSNB dataset 3 produced the FPR of 0.007 with 3792 dos instances whereas SNB dataset 2 produced a less FPR of 0.006 with 6240 instances.

It is interesting to note that class 3 or U2R of SSNB has FPR of 0.005 whereas SNB dataset 1 has 0.096 and SNB dataset 2 has 0.062. SSNB data set has only 5 U2R attack instances where as SNB dataset 1 and SNB dataset 2 has 52 attack instances each as shown in Tables I, II, and III. So even with less number of instances SSNB gives better FPR for class 3.

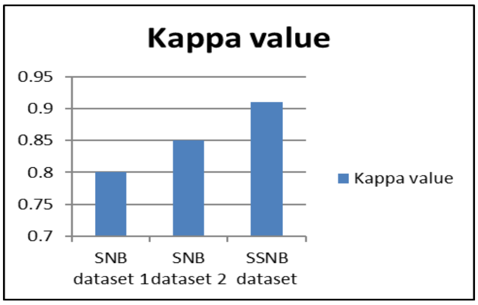

Fig. 2. Comparison of Kappa values of the 3 datasets.

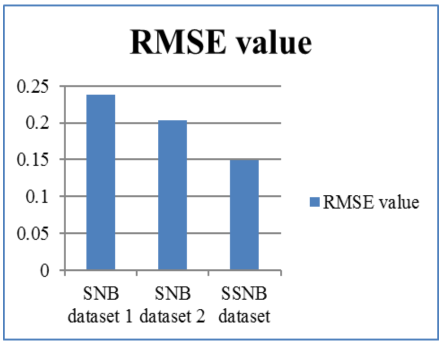

Fig. 3. Comparison of root mean squared error of the 3 data sets.

The FPR, 0.002 of SSNB for category 4 or R2L is the same as that of SNB dataset 2. But the R2L attack distribution of SSNB and SNB dataset 2 is 59 and 961 respectively.

The results obtained show that the SSNB trained using clustered KDD99 dataset has better classification efficiency when compared with that of SNB.

VI. CONCLUSION

There are many research focussed on the comparison of various classifiers using a single dataset, and analysing the effect of feature selection. But research on training dataset is minimal. The results obtained from the experiments conducted for this research suggests the possible positive impact on classification accuracy of Naive Bayes Classifier. The SSNB model achieved a classification accuracy of 94.4% where as SNB achieved 85.41% .The F-measures of SSNB are better than SNB except for probe class. This approach is thus effective in reducing the number of instances for training the Naive bayes classifier. But it has to be improved in areas such as time complexity of clustering the dataset. Other clustering algorithms can be tried in the place of k-means for better time complexity. Other classifiers can also be tested using this approach.

REFERENCES

[1] J. B. D. Cabrera, C. Gutiérrez, and R. K. Mehra, “Ensemble methods for anomaly detection and distributed intrusion detection in Mobile Ad-Hoc Networks,” *Inf. Fusion*, vol. 9, no. 1, pp. 96 – 119, 2008.

[2] S.-J. Horng, M.-Y. Su, Y.-H. Chen, T.-W. Kao, R.-J. Chen, J.-L. Lai, and C. D. Perkasa, “A novel intrusion detection system based on

hierarchical clustering and support vector machines," *Expert Syst. Appl.*, vol. 38, no. 1, pp. 306 – 313, 2011.

[3] H. G. Kayacik, A. N. Zincir-Heywood, and M. I. Heywood, "A hierarchical SOM-based intrusion detection system," *Eng. Appl. Artif. Intell.*, vol. 20, no. 4, pp. 439 – 451, 2007.

[4] S. T. Powers and J. He, "A hybrid artificial immune system and Self Organising Map for network intrusion detection," *Inf. Sci.*, vol. 178, no. 15, pp. 3024 – 3042, 2008.

[5] F. Amiri, M. R. Yousefi, C. Lucas, A. Shakery, and N. Yazdani, "Mutual information-based feature selection for intrusion detection systems," *J. Netw. Comput. Appl.*, vol. 34, no. 4, pp. 1184 – 1199, 2011.

[6] Y. Li, J. Xia, S. Zhang, J. Yan, X. Ai, and K. Dai, "An efficient intrusion detection system based on support vector machines and gradually feature removal method," *Expert Syst. Appl.*, vol. 39, no. 1, pp. 424 – 430, 2012.

[7] Y. Wei and M. Wu, "KFDA and clustering based multiclass SVM for intrusion detection," *J. China Univ. Posts Telecommun.*, vol. 15, no. 1, pp. 123 – 128, 2008.

[8] X. Gan, J. Duanmu, J. Wang, and W. Cong, "Anomaly intrusion detection based on PLS feature extraction and core vector machine," *Knowl.-Based Syst.*, vol. 40, pp. 1 – 6, 2013.

[9] C.-F. Tsai and C.-Y. Lin, "A triangle area based nearest neighbors approach to intrusion detection," *Pattern Recognit.*, vol. 43, no. 1, pp. 222 – 229, 2010.

[10] P. Casas, J. Mazel, and P. Owezarski, "Unsupervised network intrusion detection systems: detecting the unknown without knowledge," *Comput. Commun.*, vol. 35, no. 7, pp. 772 – 783, 2012.

[11] M. Tavallaee, "An adaptive hybrid intrusion detection system," The University of New Brunswick.

[12] B. Bostanci and E. Bostanci, "An evaluation of classification algorithms using Mcnemar's test," in *Proc. Seventh International Conference on Bio-Inspired Computing: Theories and Applications* (BIC-TA 2012).

Uma Subramanian is a Ph.D. candidate in the Department of Electronics and Communication Engineering of Universiti Tenaga Nasional, Malaysia. She received her B.Sc (Physics) and Master of Computer Applications degrees from Bharathidasan University, India in the years 1994 and 1997 respectively. She has worked as a lecturer for 9 years from 1998 until 2007. Her last employment was in Sunway University, Malaysia from 2003 until 2007 as a lecturer in the School of Computer Technology.

Ong Hang-See is an associate professor at College of Engineering, Universiti Tenaga Nasional (National Energy University), Malaysia. He received his Bachelor of Science degree in Electrical Engineering and Master of Science degree in Health Physics from University of North Dakota, USA, in 1986 and 1989, respectively. Later, he was educated and trained in University of Minnesota where he received his Ph. D in Biomedical Engineering and Medical Physics.

Currently, he is actively working on ICT in power utility especially in the area of smart grid. His job experience includes title like scientific and database programmer, network administrator and designer, biomedical engineer, and instrumental specialist.