Feature Reduction for Network Intrusion Detection using Principal Component Analysis in Data Mining

GERALDIN B. DELA CRUZ[1]

[1] *College of Engineering and Technology, Tarlac Agricultural University, Camiling, Tarlac, Philippines*

Abstract

Data Mining has emerged as one of the domains in the field of research. It is an analytic process designed to explore and search for consistent patterns and systematic relationships between variables in a dataset. In Data Mining, patterns in huge data are analyzed to extract useful information or knowledge. Discovering hidden information from historical data is among its important tasks while its ultimate goal is prediction. Before the data mining process, data cleaning and preprocessing are done to reduce noise and redundancy in the data. In this paper, the Principal Component Analysis (PCA) was utilized in reducing the dimensions of the KDDCup99 dataset. The goal was to reduce data dimensionality, reduce noise, and remove redundancy to find the useful feature subset that had a high influence in predicting network intrusion and reduce computational time. The study used the WEKA software in the experiment, specifically the J48, RandomTree and RandomForest decision tree algorithms, which were capable in detecting intrusions. The algorithms were first trained using 10-fold cross validation and the generated model was applied and tested. Then the results were compared over the original and reduced dataset. The results of the experiment showed improvements in detecting network intrusions in contrast to the reduced dataset over the original. This finding can be attributed to PCA as the pre- processing mechanism. It is recommended that similar studies be conducted using other classification algorithms and integrating clustering technique to perform anomaly detection and reduce the detection error rate. Future work is implementing the generated model in real time environment.

Keywords: Data Mining, Decision Trees, PCA, intrusion detection, WEKA

Introduction

Information Systems are becoming an integral part of any organizations. It is a computerized system which contains information about an organization which are useful in its various activities and functions. Computer Security is the ability to protect a computer system and its resources with respect to confidentiality, integrity, and availability. Various protocols and firewalls are in existence to protect these systems from computer threats. Intrusion is a type of cyberattack that attempts to bypass the security mechanism of a computer system. Such an attacker can be an outsider who attempts to access the system, or an insider who attempts to gain and misuse non-authorized privileges.

Meanwhile, data mining is assisting various applications for required data analysis. And it is becoming one of the techniques in intrusion detection system. Different data mining approaches like classification, clustering, association rule, and outlier detection are frequently used to analyze network data to gain intrusion related knowledge. It is an analytic process designed to explore, in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction. Predictive data mining is the most common type of data mining and one that has the most applications. The process of data mining consists of three stages: (1) the initial exploration, (2) model building or pattern identification with validation/verification, and (3) deployment.

The study described three data mining classification algorithms and how these classify intrusion detection attacks.

* **Correspondence**: Geraldin B. Dela Cruz; *Address*: College of Engineering and Technology, Tarlac Agricultural University, Camiling, Tarlac, Philippines *Email:* gbdelaccruz@tau.edu.ph;

This study was conducted to reduce the 1999 DARPA Intrusion Detection dataset using Principal Component Analysis (PCA). Specifically, it sought to: a) describe the Principal Component Analysis in reducing the features of the 1999 DARPA Intrusion Detection dataset, b) describe the following data mining algorithms, J4.8, RandomTree, and RandomForest, in detecting intrusion attacks, and c) evaluate the performance of the data mining classification algorithms on the reduced dataset in contrast to original dataset.

The results presented may serve as baseline data for other researchers when they conduct similar studies in improving and exploring the algorithms presented and for programmers and developers to develop faster and efficient intrusion detection systems.

Methodology

This study used Principal Components Analysis to reduce the features of the KDD Cup 99 dataset and decision tree induction model or simply the classification trees model, which was used to predict membership of cases or objects in the classes of a categorical dependent variable from their measurements on one or more predictor variables.

The first method was to use the 42 features from the original dataset, with the three (3) algorithms, J4.8, RandomTree, and RandomForest, in classifying the data. After the first process, using the Principal Component Analysis, the original dataset with 42 features was reduced. Using the top 10 ranked features or principal components, this reduced dataset was then classified using the same classification algorithms, results were then compared to the original dataset classification results.

The WEKA software was utilized as the data mining software in performing the processes to create the models and generated the results. The experiment followed the Input-Process-Output model.

A representative set from the original dataset was used in the experiment based on the work of Tavallaee et. al (2009). It consisted of selected records of the complete KDD 99 Cup dataset available at http://nsl.cs.unb.ca/NSL-KDD/. Although this may not be a perfect representative, the sample set did not include redundant records and the number of cases in the test set was reasonable which made it affordable to run experiments on a complete set without the need to randomly select a small portion.

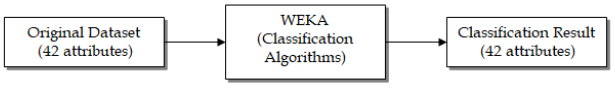

Figure 1. Procedure in classifying the original dataset

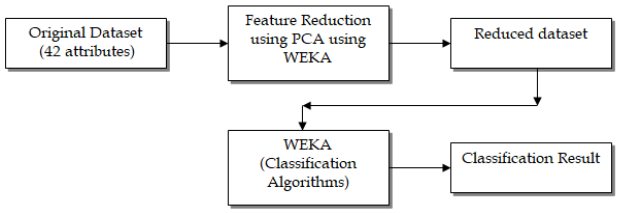

Figure 2. Procedure in reducing and evaluating the reduced dataset

The experiment followed various stages. First was to classify the original dataset using the following decision tree algorithms, J4.8, RandomTree, and RandomForest. Feature reduction followed using PCA on the original dataset with 42 attributes, to reduce the attributes. The resulting transformed dataset was then again used to classify the data using the three identified algorithms to generate the detection model. The results were compared to validate the performance of the classifiers to the reduced and original dataset.

Results and Discussion

The experiment was performed on the dataset comprising 42 attributes including the class with 11850 cases. This dataset came from the work of Tavallaee et. al (2009), which was refined and reduced to represent the whole KDDCup 99 intrusion detection dataset. The evaluation platform utilized in this study was the WEKA (Ver 3.6.10) data mining software using 10 folds cross validation. A computer with an AMD Athlon 2.8Ghz processor (eMachines LE-1660) and 1MB RAM with a 32-Bit Windows Operating System (Windows 7 SP1), was used in the experiments.

It can be seen in Table 1, that RandomForest outperformed Randomtree and J4.8 algorithms in classifying intrusions with accuracy of 97.58%, 96.80%, and 97.12%, respectively. However, Randomtree was the fastest with 0.24 seconds and with the lowest error rate. This was followed by RandomForest with 1.81 seconds and J4.8 with 2.7 seconds. However, it can be noted that the ROC was different for each of the algorithms. RandomForest had the lowest FP with 0.059, followed by RandomTree with 0.074, and J4.8 with 0.078.

Table 1. Comparison of classifier performance on original dataset with (42 attributes and 11850 cases)

	Time to build model (seconds)	% Correctly Classified	ROC Area	FP Rate	Mean Absolute Error	Relative absolute error (%)
J4.8	2.7	97.12	0.971	0.078	0.0389	13.10
RandomTree	0.24	96.80	0.95	0.074	0.0321	10.79
RandomForest	1.81	97.58	0.993	0.059	0.0349	11.74

Based on the results in Table 2, classifier accuracy was 99.27%, 99.56%, and 97.00% for RandmForest, RandomTree, and J4.8, respectively. In terms of building the model for classification, the fastest was Randomtree with 0.47 seconds, followed by J4.8 and RandonForest with 1.58 and 2.79 seconds, respectively. In terms of the FP rates and errors, the RandomTree algorithm had the lowest, followed by RandomForest and J4.8. It can also be noted that the ROC for RandomTree and RandomForest was 1 and 0.989 for J4.8.

Table 2. Comparison of classifier performance on PCA reduced dataset (10 attributes and 11850 cases)

	Time to build model	% Correctly Classified	ROC Area	FP Rate	Mean Absolute Error	Relative absolute error
J4.8	1.58	97.00	0.989	0.101	0.0458	15.42
RandomTree	0.47	99.56	1	0.001	0.0044	1.4751
RandomForest	2.79	99.27	1	0.014	0.0221	7.4428

In summary, Table 3 shows that Randomtree algorithm performed exceptionally in classifying with the PCA reduced dataset, even if it slowed down with the reduced dataset as compared with the original. Similarly, the Randomforest, improved in detecting intrusions. The experimental results showed that PCA contributed to the performance of the algorithms in performing classifying intrusions, with Randomtree and RandomForest improving its accuracy and J4.8 improving its speed. It can be noticed that all algorithms improved their performances. This implies that PCA reduced dataset improves the algorithms classification performance.

Table 3. Performance summary of the classifiers

	Original Dataset (42 attributes)			PCA Reduced Dataset (10 attributes)		
	Time	% Correctly Classified Instances	FP Rate	Time	% Correctly Classified Instances	FP Rate
J4.8	2.7	97.12	0.078	1.58	97.00	0.101
RandomTree	0.24	96.80	0.074	0.47	99.56	0.001
RandomForest	1.81	97.58	0.059	2.79	99.27	0.014

Conclusions

The Principal Component Analysis (PCA) improved the performance of the classifiers in detecting intrusions, specifically

to the simplified KDD Cup99 dataset. Using the three (3) implementations of decision trees algorithms, namely: the J4.8, Randomtree and RandomForest, classification of intrusion attacks improved by reducing the dataset using PCA. This implies that reduced datasets using PCA improves classification performance. However, the J4.8 slightly reduced classification accuracy on PCA reduced dataset but performance in time improved significantly. Conversely, both the two algorithms, RandomTree and RandomForest, improved the classification accuracy but the time was reduced.

In general, the performance of the classifiers improved on the PCA reduced dataset over the original dataset. This implies that, PCA helps in the process of detecting intrusions by reducing the dataset, removing redundancies by transforming the dataset into principal components. However, it can also negatively affect the time of detection as shown in the experiment.