International Journal of Innovation and Scientific Research
ISSN 2351-8014 Vol. 25 No. 2 Jul. 2016, pp. 586-591

http://www.ijisr.issr-journals.org/

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

Geraldin B. Dela Cruz

Institute of Engineering,
Tarlac College of Agriculture,
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, in search for consistent patterns and systematic relationships between variables in a dataset. In data mining, patterns in huge data are analyzed in order to extract useful information or knowledge. Discovering hidden information from historical data is among its important tasks and one of its ultimate goal is prediction. Prior to the data mining process, data cleaning and preprocessing is performed. In this paper, the Principal Component Analysis (PCA) was utilized to preprocess the KDD Cup 99 dataset. The goal is to address data dimensionality, by reducing noise and remove redundancy, to generate the useful feature subset that has high influence in predicting network intrusions and reduce computational time. The experiment used the WEKA software, specifically the J4.8, RandomTree and RandomForest decision tree algorithms that are capable of detecting intrusions. The algorithms was trained using ten (10) fold cross validation and the generated model was applied, tested. The results were compared between the original over the reduced dataset. Analysis of the results revealed improvements in detecting network intrusions in contrast the original dataset. This can be attributed to the PCA as a feature reduction mechanism applied as a preprocessing technique. Similar studies may be conducted using other classification algorithms and integrating other data mining techniques.

KEYWORDS: Classification, data preprocessing, decision trees, data reduction, network intrusion detection.

1 INTRODUCTION

Information Systems are becoming an integral part of organizations, which contains organizational data that serves the enterprise in its various activities and functions. These are vulnerable if not properly protected including the system and its resources with respect to confidentiality, integrity, and availability. Various protocols are in existence to protect these systems from computer threats, network intrusions and cyber-attacks that attempts to bypass the security mechanism of a computer system. Such an attacker can be an outsider attempting to access the system, or an insider who attempts to gain and misuse non-authorized privileges.

Data Mining [1] is assisting various applications [2] for required data analysis. 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, consequently, validating the findings by applying the detected patterns to new subsets of data. One of the ultimate goals 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 [3].

This study focused on the application of the Principal Component Analysis (PCA) [4] as a data preprocessing mechanism and to detect network intrusions using data mining classification algorithms. Specifically, it sought to: a) apply the Principal

Corresponding Author: Geraldin B. Dela Cruz

Component Analysis to reduce the features of the 1999 DARPA Intrusion Detection dataset. b) apply the following data mining algorithms, J4.8 [5], RandomTree [6] and RandomForest [7] to the reduced dataset in detecting intrusion attacks and c) evaluate the performance of the data mining classification algorithms to the reduced dataset in contrast to the original dataset.

The results presented herein will serve as baseline data for other researchers to conduct similar studies in exploring and improving data mining algorithms; for programmers and developers to create faster and efficient intrusion detection systems.

2 RELATED WORKS

Among the essential components in data mining is data reduction or compression technique that by applying it to the dataset, reduces the original data into smaller volume and preserves the integrity of such data. This implies that mining on reduced data is more efficient and faster while producing the same similar results. The wavelet transform and principal component analysis are implementations of the lossy data compression technique and are among the efficient methods in data reduction. The Principal Component Analysis (PCA) is similar to the Karhunen-Loeve transform which is a method for dimensionality reduction by mapping the rows of a data matrix into 2 or 3 dimensional points and that can be plotted to reveal the structure of the dataset such as in cluster analysis and linear correlations. The original data are thus projected into smaller space thus results to data compression. This technique can be utilized for dimensionality reduction. The resulting reduced dataset now composes the principal components. The first principal component explains most of the variation in the original variables [8].

There are numerous data mining algorithms introduced that can perform summarization, association, detection and other forms of data characterization and interpretation. These facilitates processing and interpretation of large data into meaningful information. The knowledge produced by data mining techniques can be represented in many different ways. It can be visualized through decision trees. Decision trees are a classic way to represent information from a machine learning algorithm, and offer a fast and powerful way to express structures in data. The J4.8 is a version of an earlier algorithm developed by J. Ross Quinlan, the very popular C4.5 [9]. The basic algorithm recursively classifies until each leaf is pure in a tree, meaning that the data has been categorized as close to perfectly as possible. This process ensures maximum accuracy on the training data, but it may create excessive rules. When tested on new data, the rules may be less effective. The overall concept is to gradually generalize a decision tree until it gains a balance of flexibility and accuracy. The algorithms can deal with both classification and regression problems.

RandomTree is a collection (ensemble) of tree predictors that is called forest**.** The classification works as follows: the random trees classifier takes the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of “votes”. In RandomTree, there is no need for any accuracy estimation procedures, such as cross-validation or bootstrap, or a separate test set to get an estimate of the training error. The error is estimated internally during the training. When the training set for the current tree is drawn by sampling with replacement, some vectors are left out. Then the classification error estimate is computed as ratio of number of misclassified vectors to all the vectors in the original data.

In the work of [10] on network intrusion detection, the RandomForest was used in the experiment with an unbalanced and balanced KDDCup99[11] datasets. Results showed that their approach provides better network detection performance by using a much smaller balanced dataset. This reduces the time to build patterns and increase detection rates. A Similar study [12] was conducted and the proposed mechanism of using RandomForest was stable and performance was well over other classification algorithms likewise it showed high detection rates.

1.1 FRAMEWORK OF THE EXPERIMENT

Fig. 1, depicts the framework of this study, which is based on the concept of Input-Process-Ouput. The model was adopted to assess the effect in terms of accuracy and speed of the algorithms in detecting intrusions, on the PCA transformed and reduced dataset compared to the original.

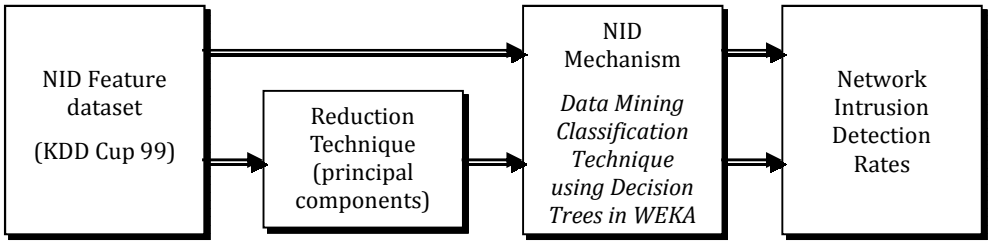


Fig. 1. The Framework of the Study

3 METHODS

Data preprocessing and classification tree analysis is one of the main techniques used in data mining. This study used principal component analysis to reduce the features of the KDD Cup '99 dataset and the three decision tree induction models or simply the classification trees model, which are 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 forty two (42) features from the original dataset, with the three (3) algorithms, J4.8, RandomTree and RandomForest to classify the cases and record the rates of detection. After the first process, using the PCA to transform the original dataset into smaller reduced feature sets called principal components. Using the top ten (10) ranked features or principal components, this reduced dataset is then classified using the three classification algorithms, results are then compared with the original dataset classification results.

The WEKA[13] data mining software was utilized in the experiment to create the models and generated the results. The experiment followed the Input-Process-Output model. The flow of the experiment is shown in Fig. 2 and Fig. 3.

A representative set from the original KDD Cup 99 dataset was used based on the study in [14]. It consists of selected records of the complete KDD 99 Cup dataset. Although this may not be a perfect representative, the sample set does not include redundant records and the number of cases in the test set is reasonable which makes it affordable to run experiments on a complete set without the need to randomly select a small portion

The experiment followed various stages. First is to train the classifiers with a representative data from the original dataset using the following decision tree algorithms, J4.8, RandomTree and RandomForest. Followed by, the application of the model generated from the training process, to asses the results of detecting intrusions from the remaining data.

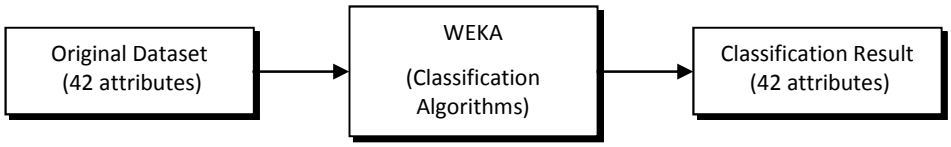


Fig. 2. Process of classifying the original dataset in WEKA

Next phase is to perform feature reduction, by applying PCA to the original dataset with forty two (42) attributes, to transform and reduce the dataset into principal components or feature sets (Fig. 3). The resulting dataset is then again used by the three identified algorithms to generate the intrusion detection model. The results are compared to validate the performance of the classifiers to the reduced over the original dataset.

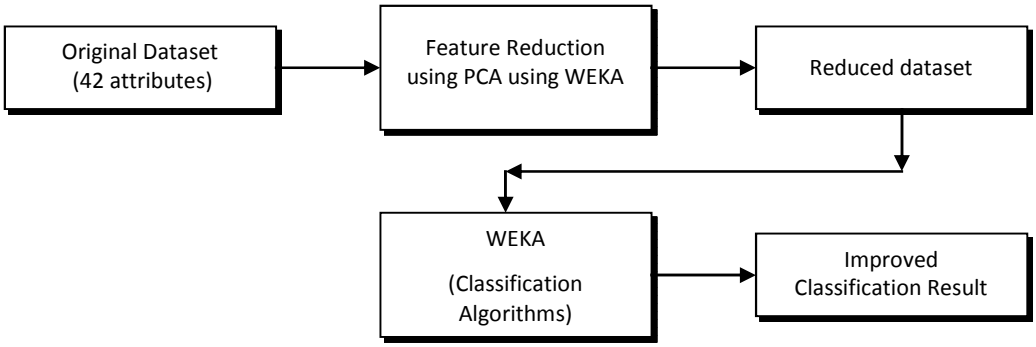
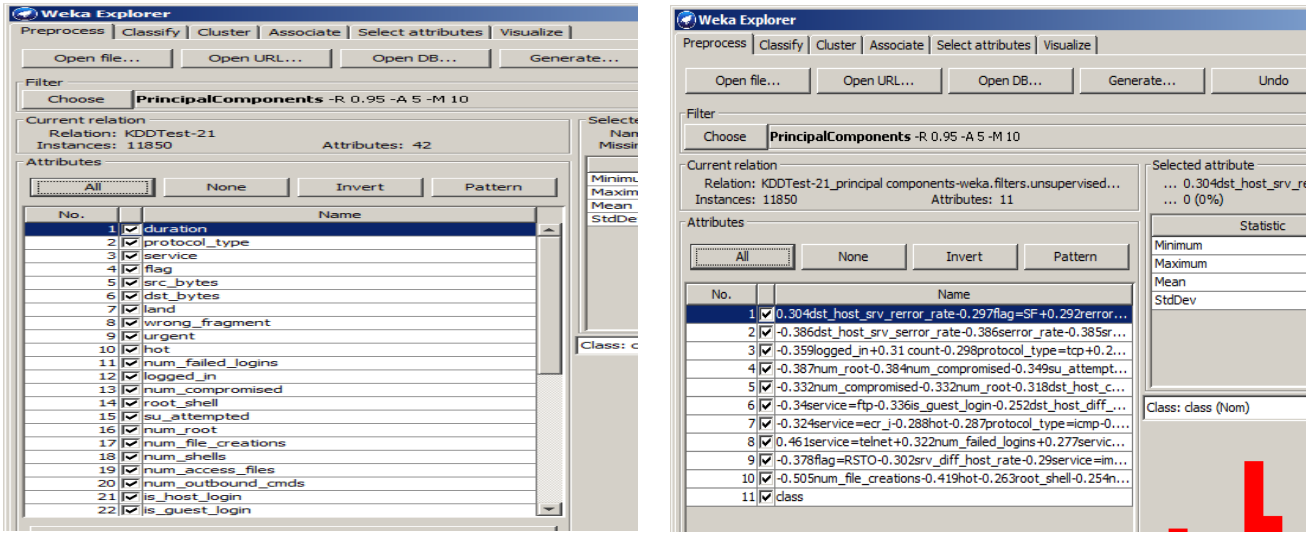


Fig. 3. Reducing the original dataset and performing classification process in WEKA

4 EXPERIMENTAL RESULTS

The experiment was performed on the dataset comprising forty two (42) attributes including the class. The transformed and reduced dataset can be seen in Fig. 4., represented by the 10 principal components. This original dataset was cleaned and reduced to represent the whole KDD Cup 99 intrusion detection dataset. The data mining software utilized in this study was the WEKA (Ver 3.6.10) running on a computer with an AMD Athlon 2.8Ghz processor, 2GB RAM with a 32 Bit Operating System.



a) b)

Fig. 4. a) The original dataset seen in WEKA. b) The transformed dataset as seen in WEKA after applying PCA

Table 1. Comparison of Classifier Performance on original dataset without feature reduction

Classifiers	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

It can be seen in Table 1, that RandomForest is more accurate than Randomtree and J4.8 algorithms in classifying intrusions with accuracy rates of 97.58%, 96.80% and 97.12% respectively. However, Randomtree was the fastest with 0.24

seconds with the lowest error rates, followed by RandomForest and J4.8 with 1.81 and 2.7 seconds respectively. Likewise, it can be noted that the ROC are different for each of the algorithms. RandomForest has the lowest False Positive rates with 0.059, followed by RandomTree and J4.8 with0.074 and 0.078.

Table 2. Comparison of Classifier Performance on PCA reduced dataset

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

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 has the lowest, followed by RandomForest and J4.8. It can also be noted that the ROC for RandomTree and RandomForest is 1 and J4.8 is 0.989.

In summary, Table 3 reveals that the RandomTree algorithm performed exceptional in classifying with the PCA reduced dataset, even processing time took longer in the reduced dataset compared with the original. Similarly, the RandomForest, improved in detecting intrusions. The experimental results show that PCA contributes to the performance of the algorithms in classifying intrusions, with RandomTree and RandomForest improving on accuracy and J4.8 improving on its processing speed. It can be noticed that performance of the algorithms has improved. This implies that a PCA reduced dataset contributes in the algorithms' classification performance. Consequently, the computational process involving intrusion detection is reduced. Thereby, proactive measures can be done, before network attacks or intrusions successful.

Table 3. Performance summary of the classifiers

Classifiers	Original Dataset (42 attributes)			PCA Reduced Dataset (10 attributes)		
	Time (seconds)	% Correctly Classified Instances	FP Rate	Time (seconds)	% 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

5 CONCLUSION

It was shown in the previous discussions that Principal Component Analysis (PCA) contributed to the improvement of the classifiers performance in detecting intrusions. Using three (3) implementations of decision tree algorithms, the J4.8, RandomTree and RandomForest, the classification of intrusion attacks significantly improves, by reducing features of the dataset using PCA, specific to the simplified KDD Cup 99 dataset. By applying PCA, the transformed and reduced datasets called principal components can be used as representative data in analyzing and detecting network intrusions, thus mining on clean and reduced dataset is more efficient.

However, the J4.8 algorithms' classification accuracy slightly dropped on PCA reduced dataset but it improved in detection time significantly. Conversely, both the two algorithms, RandomTree and RandomForest improved classification accuracy but the detection speed is negatively affected.

Generally, the performance of the classifiers accuracy improved on the PCA reduced dataset over the original dataset. Moreover, the process of detecting intrusions using decision tree algorithms using a reduced feature sets, called principal components may also introduce negative effects in the speed of detection as observed in the experiment.

FUTURE WORK

Further studies can be done, using different classifiers to explore the most fitted classifier for detection of intrusions and most compatible with the PCA data reduction method. Other methods can also be studied in enhancing and improving mechanisms for intrusion detection and prevention using other data mining techniques, and simplifying it to an anomaly detection process.

Future work is to implement the model in real time environment using real time data and the extracted classification rules to detect network intrusions, and to further validate the results of this study, specific to the PCA as a feature extractor and data reduction method. Similarly, a software agent may be developed to preprocess incoming data using PCA to further enhance the speed of detection.

ACKNOWLEDGMENT

The author would like to extend his gratitude to the Tarlac College of Agriculture in providing support in the conduct and completion of this study.

REFERENCES

[1] Witten, I. H., Frank E., *Data Mining: Practical Machine Learning Tools and Techniques*. Second Edition: Morgan Kaufmann. 2005.

[2] Dela Cruz, G. B., Gerardo, B. D., Tanguilig III, B. T., "Agricultural Crops Classification Models Based on PCA-GA Implementation in Data Mining," *International Journal of Modeling and Optimization* Vol. 4, No. 5, pp. 375-382, 2014.

[3] Han, J., Kamber M, *Data Mining: Concepts and Techniques*. Second Edition: Morgan Kaufmann, 2006.

[4] Abdi, H., & Williams, L. J., "Principal component analysis". *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433-459, 2010.

[5] Ye, Y., Wang, D., Li, T., & Ye, D., "IMDS: Intelligent malware detection system." In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1043-1047). ACM.

[6] Jacob, S. G., & Ramani, R. G., "Mining of classification patterns in clinical data through data mining algorithms." In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics* (pp. 997-1003), 2012.

[7] Breiman, L., Cutler, A., Random Forests, 2003. [Online] Available: http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm

[8] Gerardo, B. D., Lee, J., "*Principal Component Analysis Mechanism for Association Rule Mining*," School of Electronic and Information Engineering, Kunsan National University, South Korea, undated.

[9] Quinlan, J. R., *C4. 5: Programs for machine learning* (Vol. 1). Morgan Kaufmann, 1993.

[10] Zhang, J., Zulkemine, M., "*Network Intrusion Detection using Random Forests"*, School of Computing, Queen's University, Kingston, Ontario, Canada K7L 3N6, undated.
2010. [Online] Available: http://tunedit.org/repo/KDD_Cup/KDDCup99.arff

[11] Dong, S. K., Sang, M. L., Jong, S. P., "Building Lightweight Intrusion Detection System Based on Random Forest", *Advances in Neural Networks, Lecture Notes on Computer Science,* Vol, 3970, pp 224-230, 2006.

[12] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H., "The WEKA data mining software: an update". *ACM SIGKDD explorations newsletter*, *11*(1), 10-18, 2006.

[13] Tavallaee, M., Bagheri, E, Lu, W, Ghorbani, A., "A detailed Analysis of the KDD Cup 99 Data Set", Proceeedings of the 2nd IEEE Synmposium on Computational Intelligence for Security and Defense Applications, 2009.