IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

Reduced Dimensionality for Network Intrusion Detection using Principal Component Analysis

T. Manoj[1], Mr.S.Saravanakumar[2]

Research Scholar, Department of Computer Science,

Sri Jayendra Saraswathy Maha Vidyalaya College of Arts & Science, Coimbatore[1]

Head, Department of Computer Applications,

Sri Jayendra Saraswathy Maha Vidyalaya, College of Arts & Science ,Coimbatore[2]

Abstract: Intrusion Detection System (IDS) is the science of detection of malicious activity on a computer network. Due to the enormous volume existing and newly appearing network data, Data Mining classification methods are used for Intrusion Detection System. In this paper the classifying methods used are ID3, SVM, Decision Tree and One R. The data set used for this experiment is kddcup1999. The dimensionality reduction is being performed from 41 attributes to 6 and 14 attributes based on Principal Component Analysis and the 4 classifying methods are being applied. The result shows SVM method carries the highest accuracy and sensitivity with 6 and 14 attributes. J4.8 and ID3 holds the highest degree of specification for all three dimensionalities. One R has the worst Sensitivity with 6 and 14 attributes but the time taken by One R for classification is very less. It is found that the optimal algorithm may vary based on the dimensionality. Our approach focuses on using information obtained Kdd Cup 99 data set for the selection of attributes to identify the type of attack. Our work then compares the performance of the classification models by a randomly selected initial dataset with the reduced dimensionality. Furthermore, the results indicate that our approach provides more accurate results compared to the purely random one in a reasonable amount of time.

Keyword: IDS, Mining, ONER, SVM, PCA, KDD Cup99 dataset.

I. INTRODUCTION

The Internet is a worldwide network of interconnected computers enabling users to share information along multiple channels. Network Security consists of the provisions made in an underlined computer network infrastructure and policies adopted by the Network Administrator to protect the network and network accessible resources from unauthorized access, consistent and continuous monitoring and measurement of its effectiveness combined together. An intrusion detection system (IDS) is software that automates the intrusion detection process. An intrusion prevention system (IPS) is software that has all the capabilities of an intrusion detection system and can also attempt to stop possible incidents. IDS and IPS technologies offer many of the same capabilities, and administrators can usually disable prevention features in IPS products, causing them to function as IDSs. The combination of IDS and IPS known as Intrusion Detection and Prevention Systems (IDPS) is capable of detecting and preventing attacks from happening.

Data mining and IDSD

Data mining techniques can be differentiated by their different model functions and representation, preference criterion, and algorithms [17]. The main function of the model that we are interested in is classification, as normal, or malicious, or as a particular type of attack [18]. We are also interested in link and sequence analysis [12]. Additionally, data mining systems provide the means to easily perform data summarization and visualization, aiding the security analyst in identifying areas of concern [12]. The models must be represented in some form. Common representations for data mining techniques include rules, decision trees, linear and non-linear functions (including neural nets), instance-based examples, and probability models

II. REVIEW OF LITERATURE

Many of the techniques used in attempting to detect intrusion are reviewed here in this section. The most common ones are summarized below.

- Artificial Neural Networks (ANNs): Can be trained to recognize arbitrary patterns in input data, and associate such patterns with an outcome, which can be a binary indication of whether an intrusion has occurred [3].
- State Transition Tables: Describe a sequence of actions an intruder does in the form of a state transition diagram. When the behavior of the system matches those states, an intrusion is detected [4].
- Genetic Algorithms (GAs): Mimic the natural reproduction system in nature where only the fittest individuals in a generation will be reproduced in subsequent generations, after undergoing recombination and random change. The application of GAs in IDS research appeared as early as 1995, and involves evolving a signature that indicates intrusion [5]. A related technique is the Learning Classifier System (LCS), where binary rules are evolved, that collectively recognizes patterns of intrusion.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

- Bayesian Network: A set of transition rules are represented as probabilistic interdependencies in a graphical model. Each node contains the state of random variable and a conditional probability table, which determine the probabilities of the node in a state, given a state of its parent [6]. An advantage of the approach is that it can deal with incomplete data.
- Fuzzy Logic: A set of concepts and approaches designed to handle vagueness and imprecision. A set of rules can be created to describe a relationship between the input variables and the output variables, which may indicate whether an intrusion has occurred. Fuzzy logic uses membership functions to evaluate the degree of truthfulness [7].

III. PROPOSED FRAME WORK

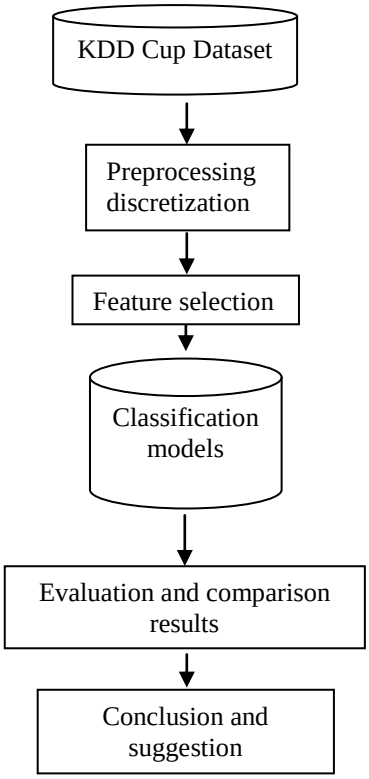


Fig. 3.1 Frame work of the Proposed Model

The data mining process of building intrusion detection models is depicted in Fig. 1.

Data Preprocessing

Normalization is used for data preprocessing, where the attribute data are scaled so as to fall within a small specified range such as -1.0 to 1.0 or 0.0 to 1.0. If using neural network back propagation algorithm for classification, normalizing the input values for each attribute measured in the training samples will help speed up the learning phase.

Dimensionality reduction

Principal component Analysis (PCA) is used for dimensionality reduction. The goal of PCA is to reduce the dimensionality of the data while retaining as much as possible of the variation present in the original dataset.

Model Evaluation using Classification Methods

In this proposed work to evaluate the performance of two different set of attributes 14 and 6 four different models of classification algorithms are used. They are namely Id3, J48, Decision Tree and SVM.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The Experimental result was conducted using KDD Cup 99 dataset with 60000 record set. The dataset consist of four different categories of attack along with normal packet dataset. The performance evaluation is based on Confusion Matrix, Accuracy, Specificity and Sensitivity. The result shows the best classification method for intrusion detection system.

Feature Selection

In machine learning and statistics, feature selection, also known as variable selection, feature reduction, attribute selection or variable subset selection, is the technique of selecting a subset of relevant features for building robust learning models. By removing most irrelevant and redundant features from the data, feature selection helps improve the performance of learning models by:

- Enhancing generalization capability.
- Speeding up learning process.
- Improving model interpretability.

Principal component analysis

PCA is a useful statistical technique that has found application in fields such as face recognition and image compression, and is a common technique for finding patterns in data of high dimension. The entire subject of statistics is based on around the idea that you have this big set of data, and you want to analyze that set terms of the relationships between the individual points in that set [4].

The goal of PCA is to reduce the dimensionality of the data while retaining as much as possible of the variation present in the original dataset. It is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences

Dataset Description

The KDD 99 intrusion detection datasets are based on the 1998 DARPA initiative, which provides designers of intrusion detection systems (IDS) with a benchmark on which to evaluate different methodologies [MIT.L.L 98]. To do so, a simulation is made of a factitious military network consisting of three ‘target’ machines running various operating systems and services. Additional three machines are then used to spoof different IP addresses to generate traffic. Finally, there is a sniffer that records all network traffic using the TCP dump format. The total simulated period is seven weeks. Normal connections are created to profile that expected in a military network and attacks fall into one of four categories:

- **Denial of Service (DoS):** Attacker tries to prevent legitimate users from using a service.
- **Remote to Local (R2L):** Attacker does not have an account on the victim machine, hence tries to gain access.

DOI 10.17148/IJARCCE.2015.41205

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

- **User to Root (U2R):** Attacker has local access to the victim machine and tries to gain super user privileges.
- **Probe:** Attacker tries to gain information about the target host.

Criteria for Evaluation*:*

To estimate the performance in the models we employed Accuracy, Sensitivity, Specificity, and ROC along with Kappa statistics, and correctly classified Instance as criteria. The accuracy, sensitivity and specificity were calculated by True Positive, False Positive, False Negative and True Negative.

Accuracy means probability that the algorithms can correctly predict positive and negative examples. Sensitivity means probability that the algorithms can correctly predict positive examples. Specificity means probability that the algorithms can correctly predict negative examples.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (3)$$

ROC curve, is a graphical plot of the sensitivity vs. (1 − specificity) for a binary classifier system as its discrimination threshold is varied

PERFORMANCE EVALUATIONS

In our work we begin with the dimensionality reduction of original dataset which consist of 41 attributes and one class label. Using Principal Component analysis obtained two set of potential dimensionalities 6 and 14 attributes. From the selected dimensionalities the experimental result shows that the performance of the reduced feature also predicts the classification in efficient manner.

Dimensionality Reduction

The original dataset consist of 41 attributes and one class label. The following will list out the attribute names

41 Attributes: duration, protocol type, service, Flag,, src_bytes, dst_bytes, land, wrong _ fragment, urgent, Hot, num_field_logins, logged_in, num_compromised, root_shell, su_attempted, num_root, num_file_creation, num_shells, num_access_files, num_outbounds_cmds, is_hist_login, is_guest_login, count, srv_count, serror_rate, srv_serror_rate, rerror _ rate, srv _ rerror_rate, same _ srv _ rate , diff _ srv _ rate , srv_diff_host_rate, dst_host_count, dst_host_srv_count, dst_hosdst_same_srv_rate,dst_host_diff_srv_rate, dst_host_same _ src _ port _ rate, dst _ host _ srv _ diff _ host _ rate, dst _ host _serror_rate, dst_host_srv_serror_rate, dst_host_rerror_rate, dst _ host_srv_rerror_rate.,

This thesis used the dimensionality reduction of original data set which is comprised of two sets of potential dimensionalities, 7 and 14 attributes are obtained by Principal Component Analysis.

Using PCA method we obtained two set of reduced dimensionalities. 7 potential attributes and 14 potential attributes which are listed as follows

6 Attributes:

flag, dst_host_diff_srv_rate , dst_host_same_srv_rate, dst_host_srv_count , srv_count , count.

14 Attributes:

flag, duration, dst_host_rerror_rate, dst_bytes,serror_rate, rerror_rate,dst_host_diff_srv_rate, dst_host_same_srv_rate,dst_host_same_src_port_rate,service,dst_host_srv_count,srv_count, count,src_bytes

Dimensionality Reduction Algorithm

- ✓ Select the dataset.
- ✓ Perform discretization for preprocessing the data.
- ✓ Apply Principal Component Analysis to filter out redundant & super flows attributes.
- ✓ Using the redundant attributes apply classification algorithm and compare their performance.
- ✓ Identify the Best One.

Comparison of Attribute Weight using Principal Component Analysis for 14 Attributes

ATTRIBUTE WEIGHT USING PCA WITH SIX ATTRIBUTES

The table 1 , 2 and 3 shows the performance of classification models with 6 , 14 and 41 attributes.

Table 1: SENSITIVITY, SPECIFICITY AND ACCURACY BASED ON 41 ATTRIBUTE FEATURE SELECTIONS

Attribute	Weight
Flag	0.91
dst_host_diff_srv_rate	0.94
dst_host_same_srv_rate	0.954
dst_host_srv_count	0.96
srv_count	0.961
Count	1

ACCURACY BASED ON 41 ATTRIBUTE FEATURE SELECTIONS

	Sensitivity	Specificity	Accuracy
SVM	97%	97%	98%
Decision Tree	95%	99%	92%
ID3	92%	99%	95%
OneR	74%	99%	86%

The table 1 shows Sensitivity, Specificity And Accuracy Based on 41 Attribute Feature Selections In Which the SVM has the highest sensitivity and accuracy of 97% and 98% respectively. Next the Decision Tree and the ID3 classified well and they produce highest degree of specificity. The worst performance is of ONER classifier.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

CHART 1: SENSITIVITY, SPECIFICITY AND ACCURACY BASED ON 41 ATTRIBUTE FEATURE SELECTIONS

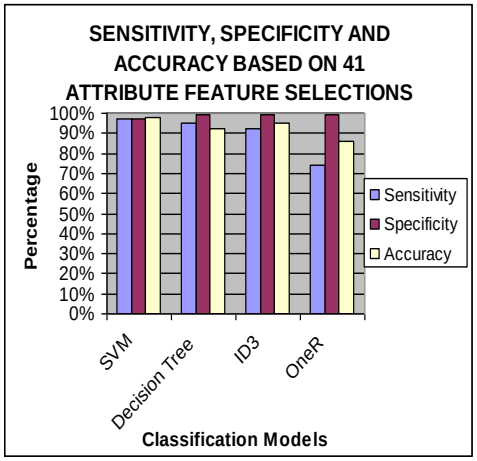
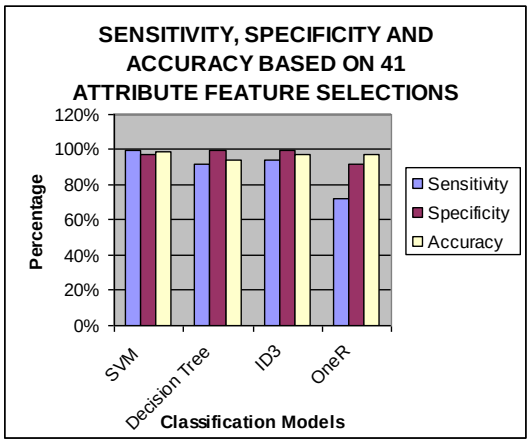


TABLE 2: SENSITIVITY, SPECIFICITY AND ACCURACY BASED ON 14 ATTRIBUTE FEATURE SELECTIONS

	Sensitivity	Specificity	Accuracy
SVM	100%	97%	99%
Decision Tree	91.5%	100%	94%
ID3	94%	100%	97%
OneR	72%	92%	97%

CHART 2: SENSITIVITY, SPECIFICITY AND ACCURACY BASED ON 14 ATTRIBUTE FEATURE SELECTIONS

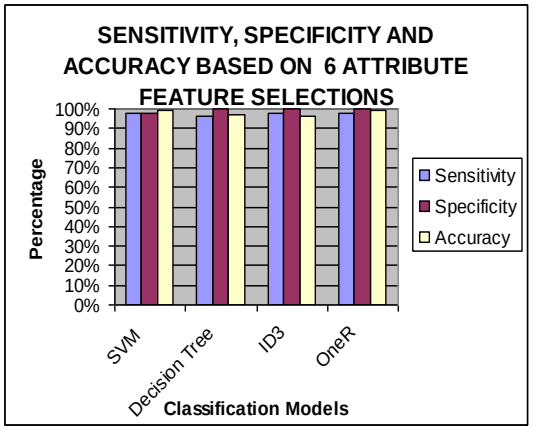


The table 2 shows Sensitivity, Specificity And Accuracy Based on 14 Attribute Feature Selections in that the SVM has the highest sensitivity and accuracy of 100% and 99% respectively. Next the Decision Tree and the ID3 classified well and they produce highest degree of specificity. The worst performance is of ONER classifier.

TABLE 3: SENSITIVITY, SPECIFICITY AND ACCURACY BASED ON 6 ATTRIBUTE FEATURE SELECTIONS

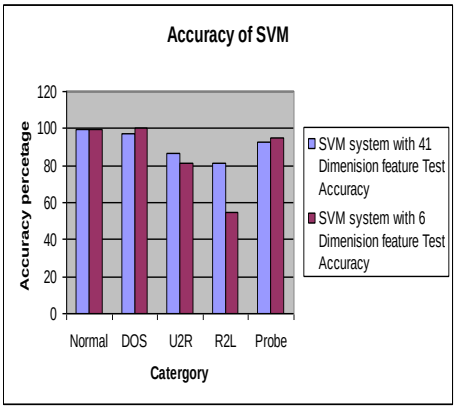
	Sensitivity	Specificity	Accuracy
SVM	98%	98%	99%
Decision Tree	96%	100%	97%
ID3	98%	100%	96%
OneR	98%	99.7%	99.5%

CHART 3: SENSITIVITY, SPECIFICITY AND ACCURACY BASED ON 6 ATTRIBUTE FEATURE SELECTIONS



The table 3 shows Sensitivity, Specificity And Accuracy Based on 6 Attribute Feature Selections that the SVM has the highest sensitivity and accuracy of 98% and 99% respectively. Next the Decision Tree and the ID3 classified well and they produce highest degree of specificity. The worst performance is of ONER classifier. On the whole the dataset with 6 attributes performs best then the remaining set of attributes

TESTING ACCURACY COMPARISON OF SVM



The above Table shows the accuracy achieved for SVMs using full dimension data (without PCA) and after the features reduction (with PCA). The testing accuracies indicate that PCA can be used to reduce data dimension without sacrificing much performance in accuracy.

V. CONCLUSION & FUTURE WORK

Intrusion Detection System is one of the major concerns in any computer networks environment. Most of the existing Intrusion Detection System uses all features in the network packet to look for known intrusive patterns. Some of these features are irrelevant or redundant Principal Component Analysis model learning algorithm is used to rank the features extracted for detecting intrusions and generate Intrusion Detection models. Results showed that the classification accuracy is increased using 6 attributes than the 41 attributes. The Time Taken by all algorithms is considerably less while using 6 attributes. Overall the J48 outperforms the remaining 3 algorithms both by using 6 & 41 attributes. Next ID3 classify well but it takes long time

DOI 10.17148/IJARCCE.2015.41205

while comparing with J48 classifier. The work can extend this experiment for identifying indeterministic type of packets.

	SVM system with 41 Dimenision feature Test Accuracy	SVM system with 6 Dimenision feature Test Accuracy
Normal	99.8	99.5
DOS	97.5	99.9
U2R	86.6	81.2
R2L	81.3	54.6
Probe	92.8	95.3

REFERENCES

[1] Axelsson S.: Intrusion Detection Systems: A Taxomomy and Survey. Technical Report No 99-15, Dept. of Computer Engineering, Chalmers University of Technology, sweden, March 2000.
[2] Bass T.: Intrusion Detection Systems Multisensor Data Fusion: Creating Cyberspace Situational Awareness. Communication of the ACM, Vol. 43, Number 1, January 2000, pp. 99-105.
[3] Debar H., Dacier M., Wespi A.: Towards a taxonomy of intrusion-detection systems. Computer Networks, 31, 1999, pp. 805-822.
[4] Dorosz P., Kazienko P.: Omijanie intrusion detection systems. Software 2.0 no 9 (93), September 2002, pages 48-54. (In Polish only).
[5] Dorosz P., Kazienko P. Systems wykrywania intruzów. VI Krajowa Konferencja Zastosowan Kryptografii ENIGMA 2002, Warsaw 14-17 May 2002 , p. TIV 47-78,
Elson D: Intrusion Detection, Theory and Practice. March 27, 2000.
[6] Fan W., Miller M., Stolfo S., Lee W., Chan P.: Using Artificial Anomalies to Detect Unknown and Known Network Intrusions. In Proceedings of the First IEEE International Conference on Data Mining, San Jose, CA, November 2001.
[7] Frederick K. K.: Network Intrusion Detection Signatures. December 19, 2001.
[8] Intrusion Detection Systems (IDS). Group Test (Edition 3), NSS Group, July 2002.
[9] Jones A.K., Sielken R.S.: Computer system intrusion detection: a survey. 09.02.2000.
[10] Lee W. i inni: A data mining and CIDF based approach for detecting novel and distributed intrusions. Recent Advances in Intrusion Detection, Third International Workshop, RAID 2000, Toulouse, France, October 2-4, 2000, Proceedings. Lecture Notes in Computer Science 1907 Springer, 2000.
[11] Lee W., Stolfo S, Mok K.: Adaptive Intrusion Detection: a Data Mining Approach. Artificial Intelligence Review, 14(6), December 2000.
[12] Manganaris S., Christensen M., Zerkle D., Hermiz K.: A data mining analysis of RTID alarms. Computer Networks, 34, 2000, pp. 571-577.
[13] Ajit Abraham, Ravi Jain, Johnson Thomas, Sang Yang Han " D-SCIDS: Distributed softcomputing intrusion detection system" Journal of Network and Computer Applications, Elsevier, 2005.
[14] Susan M. Bridges and M. Vaughn Rayford, "Fuzzy data mining and genetic algorithms applied to intrusion detection," in Proceedings of the Twenty-third National Information Systems Security Conference. National Institute of Standards and Technology, Oct. 2000.
[15] T.S.Chou, K.K. Yen and J.Luo " Network Intrusion Detection Design Using Feature Selection of Soft Computing Paradigms" International Journal of Computaional Intelligence Vol. 4 Number 3,2007
[16] J. Gomez and D. Dasgupta, "Evolving fuzzy classifiers for intrusion detection," in Proceedings of the 2002 IEEE Workshop on the Information Assurance, West Point, NY, USA, June 2001
[17] S. B. Cho, "Incorporating soft computing techniques into a probabilistic intrusion detection system," IEEE transactions on systems, man and cybernetics, application and reviews, vol.32, pp.154-160, May 2002.
[18] D. Bulatovic and D. Velasevic, "A distributed intrusion detection system based on bayesian alarm networks," Lecture Notes in Computer Science (Secure Networking CQRE (Secure) 1999), vol. 1740, pp. 219–228, 1999.
[19] D. Barbara, N. Wu, and S. Jajodia, "Detecting novel network intrusions using bayes estimators," in Proceedings of the First SIAM International Conference on Data Mining (SDM 2001), Chicago, USA, Apr.2001.
[20] M. Bilodeau and D. Brenner, Theory of multivariate statistics. Springer - Verlag : New York, 1999.Electronic edition at ebrary, Inc.
[21] WEKA: Data Mining Software in java (2008).

DOI 10.17148/IJARCCE.2015.41205