International Journal of Security and Its Applications
Vol. 10, No. 8 (2016), pp.151-158
http://dx.doi.org/10.14257/ijsia.2016.10.8.13

A Fusion of Feature Extraction and Feature Selection Technique for Network Intrusion Detection

Yasir Hamid[1], M.Sugumaran[2] and Ludovic Journaux[3]

[1]*Research Scholar, Dept. of CSE Pondicherry Engineering College*
[2]*Professor and Head, Dept. of CSE, Pondicherry Engineering College*
[3]*Associate Professor, University of Burgundy*
[1]*bhatyasirhamid@pec.edu*,[2]*sugu@pec.edu*,[3]*ljourn@gmail.com*

Abstract

With varied and widespread attacks on information systems, intrusion detection systems (IDS) have become an indispensable part of security policy for protecting data. IDS monitor event logs and network traffic to uncover suspicious connections that deviate from the regular profile and identify them as threats or attacks. Like most of the cases the dataset used for intrusion detection i.e., KDD99 suffers two problems: imbalanced class distribution and curse of dimensionality. In this work SMOTE has been used for balancing the dataset and once balanced, Principal Component Analysis (PCA) has been used to extract the features. And after that on the transformed dataset Correlation based Feature Selection (CFS) is used to select a subset of important features. The reduced dimension dataset is tested with Support Vector Machines (SVM). Obtained results demonstrate improved detection accuracy, computational efficiency with minimal false alarms and less system resources utilization

Keywords*: Intrusion Detection, Machine Learning, PCA, SVM*

1. Introduction

The popularity of Internet in fields of business, entertainment, education and science is ever increasing, making it most important means of communications around. Being the indispensible means of communication of modern era, it should be safeguard from the users with destructive mindset *i.e.,* hackers. Their full time job is to compromise the security mechanism of the site. They launch a set of attacks on the system to compromise the security of network system. No matter how cautiously a network is laid, secure is always a relative term, given enough time and resources, the attackers will eventually find a way to break into the system [1]. To cater the security issues, firewalls have been used for long time now. Firewall is able to scrutinize incoming and outgoing traffic and block the traffic as per written policy. Even if it detects an attack, generation of an alert is not a property of the firewall. An attack is any activity that compromises the Confidentiality, Availability and Integrity of the system [2] .One such group of attacks that user tries to attack any security parameter is called intrusion, and care should be taken to protect the system and its resources from being taken over by the attacker. The process of safeguarding the system from malicious users and warranting it's availability to the qualified users is called intrusion detection and the device against whom the responsibility of securing the system and its resources lies, is called intrusion detection system [3].

With booming number of intruders and hackers in today's vast and sophisticated computerized world, it is increasingly challenging to identify unknown attacks in promising time with no false positive and no false negative [4]. Based on the modeling approach and detection techniques intrusion detection systems are classified into two major types, misuse based detection and anomaly based detection. With the former

ISSN: 1738-9976 IJSIA

dealing with known attack patterns, which are known to the IDS while the later dealing with both known and unknown attacks. Major concern with misuse detection is its inefficiency to detect new attacks without the corresponding signature in the database. So new attacks can successfully bypass the misuse detection based approach which results in system damage. Anomaly based detection deals with both known and unknown attacks and its detection efficiency is more compared to misuse detection based approach due to new attack detection capability. Anomaly based intrusion detection approach detects network behavior either as normal or abnormal and it is inefficient to spot exact type of attack. Major concern in anomaly based intrusion detection is generation of false alarms due to improper design of detection approach. Based on the scope of surveillance IDS can be Host-Based or Network-Based. Network based IDS analyze the network packets from network's ongoing traffic. After monitoring network traffic these systems try to detect abnormal activities. If system finds the intrusions, it sends the alert to the administrator of particular computer network. Contrary to this Host based Intrusion Detection System based on analyzing the actions which are happening on a particular computer. It collects the suitable information from the activities of the host and it uses this information for finding intrusions [5].

For last two decades a lot of research has been done on intrusion detection using machine learning techniques, all the research efforts documents have been carried offline on some dataset as all the classification technique need labeled data for training them, which is not possible in case of real networks. Hence researchers have tested their approaches against the already collected and labeled dataset in the offline mode, most of them have taken KDD99 cup dataset in consideration. The dataset is imbalanced one with 42 attributes and suffers curse of dimensionality. The Curse of dimensionality is a fancy term that groups the set of problems encountered while analyzing the high dimensional dataset, failing to visualize the dataset, more resources needed and lesser accuracy being some [6]. Seldom are all the dimensions of the dataset needed for effective categorization of the data. Dimensionality reduction is the set of techniques that transform the data from high dimensions space to low dimensional space thereby reducing the complexity of the problem to be solved. In this paper a feature extraction technique PCA for dimensionality reduction of the dataset has been used.

Not all the dimensions of dataset are equally important for classification and hence same accuracy can be achieved with lesser dimensions. Dimension reduction is the process of transforming high dimensional dataset into lower dimensionality space [7]. Dimension reduction techniques can be based on feature selection or feature extraction. The feature selection techniques take on high dimensional dataset and select a subset of attributes of the dataset based on some discriminating ability of the dataset. While as feature extraction techniques use linear or no linear projections to project the dataset with the projecting space being formed by some calculations of the original attributes. Feature selection techniques may be laid as a filter wrapper or hybrid form, depending on whether the selection process of the technique is driven by intrinsic dimensionality of the dataset or the classifier output [8]. For a particular classifier wrapper approaches select the best features but need lot of computational resources than the filter approaches which select the attributes based on intrinsic dimensionality rather the classifier.

In this paper Principal Component Analysis is used for feature extraction. PCA is a linear projection technique that takes on n-dimensional dataset and transforms it in some other space where the attributes are ordered on the basis of variance. Greater the variance of the attributes higher is the importance of the attribute in discriminating between different classes. PCA can be used to reduce the dimensionality of dataset while still preserving most of the variance by creating a new feature space based on the top *k* eigenvectors (according to their eigen values). However, the "new" features will be different from the original one.

PCA transforms the dataset different space and this need not be necessarily having lower dimensions and hence a transformed feature space should be subjected to feature selection[9]. Feature selection is the set of techniques that reduce the complexity of the dataset by eliminating the redundant dimensions of the dataset. In this work, **correlation based feature selection** (CFS for feature selection. CFS focuses on selecting a set of attributes from the dataset that are highly correlated to the class and have very little co-relation among themselves. Once the prepossessing step of the dataset is over, the reduced dataset is fed to SVM that finds a hyper-parameter in higher dimensional space for discriminating the normal data from the attacks. Various performance metrics taken into consideration to compare different approaches are: TP rate, FP rate, Precision, Recall, ROC area, F-measure.

The rest of the paper is organized as follows. In Section 2 a review of the related works in network intrusion detection and machine learning is given. Overview of SVM, PCA technique and other techniques is given in Section 3. Section 4 discusses about the experimental setup of this work. Results are presented in Section 5, finally Section 6 concludes the paper.

2. Literature Survey

A complete survey of intrusion detection attempts using machine learning is given in our work [10], from the paper it is concluded that SVM is one of the most popular techniques for designing an intrusion detection system. Authors in [11] have attempted to compare the SVM with neural networks, even thought the accuracy of the two is more or less the same , time taken to build up the model is less is less in SVM. In [12] the researchers fused the SVM's with clustering approach, firstly the clustering was done on data then SVM was developed for each different cluster, this approach had comparative lesser detection rate for U2R and R2L attacks. Authors in [13] have attempted to study the viability of eliminating some of the attributes of the dataset without compromising on the detection rate and had used CFS for feature selection. The detection rate of the reduced dataset was rather improved as CFS selected the best features and eliminated the redundant features. In [14] the attempt was made to compare CFS with other wrapper based approaches and results show that CFS compares favorably with wrapper approach but requires much less computational resources. In [15] PCA was used for feature extraction and SVM was used as a classifier , results show that the mentioned approach outperformed all its competitors in terms of detection rate and computational time. Authors in [16] have used over-sampling and under-sampling technique to balance the dataset before using RIPPER classifier to build a model for it, results overwhelmingly favor the balanced dataset over the imbalanced dataset. To counter the problem of skewed class distribution in KDD99 dataset the authors in [17] have used SMOTE to oversample the minority class instances, once the data is balanced Information Gain was used to select the subset of all the features and results showed that such approach has higher prediction rate than its archrivals. A distance based feature extraction was reported in [18] with the n-dimensional dataset being transformed to a single dimensional distance space the proposed model had lesser detection rate for U2R and R2L attack groups.

3. Techniques Used in Work

This section discusses about various techniques and methods used in this work. A clear explanation about all the techniques used in this work is presented

3.1. Support Vector Machine (SVM)

SVM based on the hyper-plane, are learning systems that use hypothesis space of linear function in a high dimensional feature space. Due to the popularity and widespread use,

reporting satisfactory results in other classification domains in this work SVM has been used as a base classifier. SVM is based on the hyper-plane and transforms the data from lower dimension into the higher dimensions and there in high dimensional space tries to find a hyper-plane that effectively distinguishes two or more classes. From work [19] Given n training data points $\{(x_1, y_1), (x_2,y_2), (x_3, y_3), ..., (x_n , y_n)\}$, where xi $\in R^d$ and $y_i \in \{+1, -1\}$. Consider a hyper-plane defined by (w, b), where w is a weight vector and b is a bias, new object *x* can be classified with

$$f(x) = sign(\sum_i^n \alpha_i y_i (x_i.x) + b) \text{----------} \quad (1)$$

3.2. Principal Component Analysis

Principal Component Analysis (PCA) invented by by Karl Pearson, widely used in feature extraction and multivariate analysis, uses an orthogonal transformation for converting a set of observations of possibly correlated variables into a set of values of uncorrelated variables called principal components with the first component having maximum variance[9], with the number of principal components being less or equal to original features. The popularity of PCA comes from three important properties. First, it is the optimal (in terms of mean squared error) linear scheme for compressing a set of high dimensional vectors into a set of lower dimensional vectors and then reconstructing the original set. Second, the model parameters can be computed directly from the data - for example by diagonalzing the sample covariance matrix. Third, compression and decompression are easy operations to perform given the model parameters - they require only matrix multiplication. The PCA model can be represented by

$$u_{mx1} = W_{mxd} X_{dx1} \text{ ----------} \quad (2)$$

Where *u,* a m-dimensional vector is a projection of x- the original d-dimensional vector (m<<d). PCA takes on the data and transforms the data by some linear combination of the attributes. What makes PCA different from feature selection algorithm is the case that feature selection algorithms select a subset of raw features while as PCA extracts the features by some linear combination. PCA orders the transformed data according to the variance as the variance determines the effectiveness of the attribute in the process of classification. Once ordered based on the classification the user can select the topmost principal components and use them for building a classifier.

3.3. Correlation Based Feature Selection

Principal Component Analysis transforms the data into the principal components and orders the components based on the variance. Greater the variance of the component important the feature is for the purpose of classification. Not all the components are necessary for the purpose of classification, hence Correlation based Feature Selection (CFS) is used to select the subset of the features. CFS uses a search algorithm along with the function to evaluate the merit of the feature subsets [13]. An ideal feature is one that is highly correlated with the class variable and has minimum correlation with other features. To measure the correlations between features and the class, and between features a measure based on conditional entropy is used. For two variables random X and Y drawn from the ranges R_x and R_y. The entropy of Y before observing X is given in equation3.

$$H(Y) = -\sum_{y \in R_y} p(y) \log(p(y)) \text{----------} \quad (3)$$

The entropy of Y after observing X is given by equation 4

$$H(Y) = -\sum_{x \in R_x} P(X) \sum_{y \in R_y} p(y|x) \log(p(y|x)) \text{----------} \quad (4)$$

The correlation of Y on X is given in equation 5

$$H(Y|X) = \frac{H(Y)-H(Y|X)}{H(Y)}\text{----------} \quad (5)$$

3.4. Synthetic Minority Oversampling Technique(SMOTE)

Classifier learning with data-sets that suffer from imbalanced class distributions is a challenging problem in data mining community. Imbalanced class problem adds up the complexity to the classification process and most often the classifier tends to be biased towards the less frequent classes of data. Learning from such data sets that contain very few instances of the minority (or interesting) class usually produces biased classifiers that have a higher predictive accuracy over the majority class (es), but poorer predictive accuracy over the minority class. Researchers have dealt with imbalanced class problem by re-sampling (over sampling or under sampling) or synthesizing the infrequent classes. In this paper a technique **Synthetic Minority Oversampling Technique (SMOTE)**[20] has been used. Developed by **Nitesh V. Chawla,** SMOTE takes up a dataset and tries to balance the dataset by synthetically producing the instances of the minority classes. SMOTE is different from other oversampling techniques in the case that SMOTE doesn't replicate the instances rather it produces the similar instances. Simple replication of few instances results in over fitting of the model and a more realistic approach is to produce instances similar to existing instances without replicating the instances. The values for the new instances are calculated by applying some mathematical formula on the existing data records. For each item a set of k nearest neighbors is selected and new instances are fabricated as given in equation 5.

$$New.Feat = Cur.Feats + (Cur.Feats - Nbr.Feats) * \ Rnd(0,1) \quad \text{-------} \quad (5)$$

4. Experimental Setup

The pictorial representation of proposed work is given in Figure 1. The model consists of a series of modules to be followed in sequential order. The process starts with the normalization of raw data in the range 0-1. Once normalized, the data is fed to SMOTE which finds out the minority classes in the data and synthetically produces new instances. Not only is the raw data oversampled but also the major classes are under sampled as the data after even after oversampling was highly imbalanced. Once dataset is balanced, PCA is implemented on it, to transform the dataset into principal components ordered by variance. PCA just projects the data into some other space, it does not actually reduce the features of the dataset. Not all the components of the projected dataset are equally needed for the purpose of classification. So to select a subset of components of the dataset Correlation based feature selection (CFS) technique is used. CFS selects a subset of

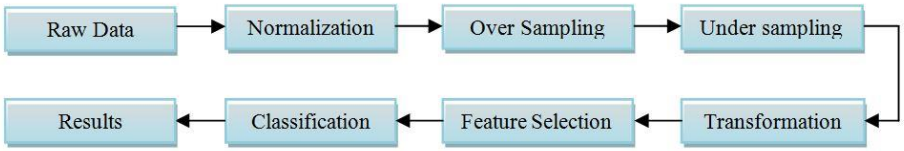

Figure 1. Experimental Setup

features from the dataset that are highly correlated to the class variable and least correlated with other features. This reduced feature set is fed to SVM with the gamma value of 0.0, epsilon of 0.001 and the kernel type set to radial bias the to build up the classifier. For the purpose of evaluation 10 fold cross validation has been used. All the

experiments were carried on dell i3 3.60GHz computer with 4GB Ram. Weka tool a free to use repository of machine learning techniques written in java has been used for the experiments.

5. Results and Discussions

Three different datasets have been used in this work. The first one consists of randomly selected 100000 records with 20219 and 79781 attacks spanning across 5 groups as shown in **Table 1** given below.

Table 1.

SNO	Dataset	Normal	DOS	PROBE	U2R	R2L	Total
1	**Raw Dataset**	20219	78635	830	30	286	100000
2	**Over Sampled**	20219	84031	9044	6480	14103	133887
3	**Over & Under Sampled**	20216	18338	9044	6480	13508	67586

This dataset is actually imbalanced and is fed to **SMOTE**, SMOTE synthetically produces the instances of the minority classes and results in a dataset of 133887 instances with 20219 normal instances and 113668 being attacks. Not only have we oversampled the minority classes but in a different set we have under-sampled the majority by using random under-sampling class resulting in a dataset containing of 67586 instances with 20216 normal and 47370 being attacks.

All the three dataset have been normalized and fed to the SVM using 10 fold cross validation and the results are documented in the Table 2 given below.

Table 2.

SNO	Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
1	**Raw Dataset**	0.991	0.002	0.983	0.991	0.986	0.994
2	**Over Sampled**	0.974	0.002	0.976	0.974	0.97	0.986
3	**Over & Under Sampled**	0.948	0.009	0.953	0.948	0.939	0.969

In the second phase of experiment PCA is used to transform all the dataset, PCA uses linear mathematical formula to transform the data. These transformed datasets fed to SVM with default configuration and the results are documents in **Table 3**. In this phase of experiments all the features of the transformed dataset are used for the purpose of classification.

Table 3.

SNO	Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
1	**Raw Dataset**	0.998	0.002	0.998	0.998	0.998	0.996
2	**Over Sampled**	0.996	0.002	0.996	0.996	0.996	0.998
3	**Over & Under Sampled**	0.998	0.002	0.99	0.998	0.989	0.994

Not all the attributes of the transformed are equally important for the purpose of classification; hence CFS was used to reduce the dimensions. Table 4 given below lists out the attributes of datasets spared in the dataset after feature reduction.

Table 4.

SNO	Dataset	TP Rate FP Rate	Selected Attributes
1	**Raw Dataset**	81	1,2,68 : 3
2	**Over Sampled**	77	1,2,6,12,52,55,56,65 : 8
3	**Over & Under Sampled**	79	1,2,3,5,6,7,10,11,15,26,27,30,31,33,49,50,62,63,64 : 19

These reduced are fed to SVM. The SVM configurations are kept same throughout. The results of SVM are listed in Table 5 below.

Table 5.

SNO	Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
1	**Raw Dataset**	0.992	0.002	0.985	0.99	0.986	0.994
2	**Over Sampled**	0.984	0.002	0.984	0.984	0.984	0.992
3	**Over & Under Sampled**	**0.999**	**0.001**	**0.993**	**0.997**	**0.993**	**0.994**

As it is evident from the table, 5 that SVM performs better on the balanced dataset.

6. Conclusion

In this work we have attempted to minimize the effects of imbalance class problem on the performance metrics. We have focused on intrusion detection dataset KDD99 dataset which is inherently imbalanced. The SMOTE algorithm was used to balance the imbalanced dataset. After balancing the dataset all the attributes were normalized to 0-1 range. This balanced and normalized dataset was operated on by PCA. PCA transformed the dataset into principal components ordered by variance. Correlation bases Subset Evaluation feature selection was used to select the most important attributes of the dataset. For the purpose of classification we have used SVM. Results documented show that our approach has improved performance in terms of precision, recall and ROC than all its counterparts.

References

[1] D. Challener, K. Yoder, R. Catherman, D. Safford and L. Van Doorn, "A practical guide to trusted computing", Pearson Education, **(2007)**.
[2] R. J. Ellison, D. A. Fisher, R. C. Linger, H. F. Lipson and T. Longstaff, "Survivable network systems: An emerging discipline", DTIC Document, **(1997)**.
[3] D. J. Hacherl, P. Garg, M. D. Satagopan and R. P. Reichel, System and method for protecting domain data against unauthorized modification. Google Patents, **(2007)**.
[4] D. L. Shinder and M. Cross, Scene of the Cybercrime. Syngress, **(2008)**.
[5] S. Axelsson, "Intrusion detection systems: A survey and taxonomy", Technical report Chalmers University of Technology, Goteborg, Sweden, **(2000)**.
[6] J. A. Lee and M. Verleysen, Nonlinear dimensionality reduction. Springer Science & Business Media, **(2007)**.
[7] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection", J. Mach. Learn. Res., vol. 3, **(2003)**, pp. 1157–1182.
[8] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning", Artif. Intell., vol. 97, no. 1, **(1997)**, pp. 245–271.
[9] I. Jolliffe, Principal component analysis. Wiley Online Library, **(2002)**.
[10] Y. Hamid, M. Sugumaran and V. Balasaraswathi, "IDS Using Machine Learning - Current State of Art and Future Directions", Br. J. Appl. Sci. Technol., vol. 15, no. 3, **(2016)** Jan., pp. 1–22.
[11] S. Mukkamala, G. Janoski and A. Sung, "Intrusion detection using neural networks and support vector machines", in Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on, vol. 2, **(2002)**, pp. 1702–1707.
[12] 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, **(2011)**, pp. 306–313.
[13] M. A. Hall, "Correlation-based feature selection for machine learning", The University of Waikato, **(1999)**.
[14] M. A. Hall and L. A. Smith, "Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper", in FLAIRS conference, vol. 1999, **(1999)**, pp. 235–239.
[15] X. Xu and X. Wang, "An adaptive network intrusion detection method based on PCA and support vector machines", in Advanced Data Mining and Applications, Springer, **(2005)**, pp. 696–703.
[16] D. A. Cieslak, N. V. Chawla and A. Striegel, "Combating imbalance in network intrusion datasets", in GrC, **(2006)**, pp. 732–737.

[17] A. Tesfahun and D. L. Bhaskari, "Intrusion detection using random forests classifier with SMOTE and feature reduction", in Cloud & Ubiquitous Computing & Emerging Technologies (CUBE), 2013 International Conference on, **(2013)**, pp. 127–132.
[18] W.-C. Lin, S.-W. Ke and C.-F. Tsai, "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors," Knowl.-Based Syst., vol. 78, **(2015)**, pp. 13–21.
[19] P. Manandhar, "A Practical Approach to Anomaly-based Intrusion Detection System by Outlier Mining in Network Traffic", Masdar Institute of Science and Technology, **(2014)**.
[20] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," J. Artif. Intell. Res., **(2002)**, pp. 321–357.

