International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 12 (2017) pp. 3370-3376

Evaluation of PCA and K-means Algorithm for Efficient Intrusion Detection

N. Chandra Sekhar Reddy [1]
Professor, Dept. of CSE, MLR Institute of Technology, Hyderabad, India.

Purna Chandra Rao Vemuri[2]
Professor, Dept. of CSE, Swamy Vivekananda Institute of Technology, Hyderabad, India.

A. Govardhan[3]
Professor, Dept. of CSE, Principal, JNTUH , Hyderabad, India.

[1]ORCID: 0000-0002-8699-176X, Researcher ID: Q-1485-2016

Abstract

The prevention of intrusion in systems is guaranteed and an intrusion detection system is amazingly attractive in detecting the fraud with powerful intrusion detection system. Extreme work is done on intrusion detection systems, however at the same time these are not tremendous because of high number of false alerts. One of the main sources of false alerts is because of the utilization of a crude dataset that contains repetition. Recently, principal component analysis (PCA) has been utilized for feature extraction and in which components are fundamentally anticipated into a principal space and at that point elements are chosen based on the Eigen values, yet the elements with the largest Eigen values might not have the certification to give ideal affectability to the classifier. To keep away from this issue, an improvement technique is required. In this paper, it is proposed that IDS with the union of best efficient features selected by Principal Component Analysis (PCA) can reduce the computational complexity of the system. Along these lines, we used an improved version of K-means clustering algorithm for classification and enhancing the accuracy. We experimented this approach on intrusion detection KDDCup'99 benchmark dataset obtained from UCI Machine learning repository.

Keywords: Intrusion detection, PCA feature extraction, k-means, Enhanced k means classification, and KDDCup'99 Data Set. component; formatting; style; styling; insert (key words)

INTRODUCTION

Clearly organizations for doing their day by day business rely on upon Internet and PC systems, in this manner shielding their business from potential assaults or anomalous exercises went for trading off their systems should be deliberately concerned [1-2]. Distinctive frameworks, for example, firewall, client verification, antivirus and information encryption intended to secure PCs systems [2], however these conventional system intrusion frameworks have neglected to totally ensure systems as a result of expanding and modern nature of new assaults [3]. IDS was developed as a tool for detecting attacks mounted over the network [19].

Along these lines, intrusion detection system (IDS) turn into an indistinguishable part of every PC systems to screens and investigations arrange traffics to distinguishes the dangers, assaults and anomalous occasions before they harm associations' important data resources [1-4]. An intrusion detection framework comprise of information gathering, information clearing and pre-preparing, recognizing the intrusion, reporting and do a sensible activity which in these procedure identifying the assault is a fundamental part [4]. As per[5], IDS characterized as a product that have a capacity to computerize the intrusion detection. High order precision and a low false alert rate are the two primary attributes of very much created IDS [1].

Rest of the paper is organized as follows: Section 2 provides the overview of related work. Section 3 presents the theoretical background related to PCA and K-means clustering algorithms. Section 4 describes the proposed methodology. Section 5 gives performance metrics, experimental results, and discussions. Finally, conclusion is given in Section 6.

RELATED WORK

As a rule, IDSs are sorted into two fundamental gathering in light of their detection approaches: oddity and abuse (signature) detection [2-4]. In inconsistency detection frameworks, it assume that anomalous conduct is exceptional and not the same as would be expected conduct, so any

deviations from ordinary exercises considered as a nosy conduct and the model will be manufacture in light of the typical information [1-4]. Then again, abused detection framework denoted the review information as an intrusion on the off chance that it is coordinated with the predefined mark of assaults [4]. Both of these methodologies have a few upsides and downsides. Irregularity detection frameworks has better execution in recognizing the already obscure assaults, additionally has high false alert rates, since any deviation from ordinary exercises named strange conduct [1]. Abuse detection frameworks, however has a low false alert rate yet fall flat when confronting already obscure assaults [1-4].

Albeit numerous endeavors has been done on building a viable detection frameworks, all things considered it is likewise an open research range. Since intrusion detection issue has diverse perspective, for example, expansive size of movement information, unequal information sets, questionable limits amongst ordinary and meddlesome conduct and exceptionally dynamic environment, so every specialist address some of these perspectives [4]. Support Vector Machine (SVM) as well-known grouping methods has been utilized to enhance the precision of IDS.

Authors utilized SVM to enhance intrusion framework in remote neighborhood [6]. To enhance the execution of SVM, [7] utilized both SVM and Artificial Neural Networks (ANN). Be that as it may, to improve low-regular assault's detection and detection security of ANN-based IDS, [8] proposed an approach called FC-ANN, in light of ANN and fluffy grouping. Most as of late, an IDS was presented by coordinating insightful element swarm based harsh set (IDS-RS) for highlight choice and streamlined swarm improvement for arrangement [3][23]. As state in [4], use of computational knowledge (CI) in intrusion detection frameworks pulled in the consideration of many examines to itself. CI techniques, for example, manufactured invulnerable frameworks, fake neural systems, swarm insight and delicate processing demonstrated a better execution contrast and the conventional strategies in high computational speed, adaptation to non-critical failure and taking care of uproarious information sets. Since as of late could processing has been tended to by scientists, [5] give an exhaustive audit on intrusion detection and prevention systems (IDPS) in distributed computing frameworks. As proposed in [9], coordinating swarm insight, particularly molecule swarm streamlining with the other machine learning classifier to diminish the preparation time and enhance the productivity of IDS is an opening examination zone.

THEORETICAL BACKGROUND

The following subsections provide the necessary background to understand the problem.

Principle Component Analysis (PCA) :

PCA is a statistical method normally used for data analysis and is a very useful method of feature selection. The PCA is applied to transform raw features into principal features so that the features are more clearly visible and their importance is visualized. This technique has been used from last few years in different domains. In this technique, the features are selected on the basis of Eigen values, the features with higher Eigen values are selected and the features with lower Eigen values are ignored.

Research on various dimensionality reduction techniques rationalize that Principal Component Analysis is a proved technique for dimensionality reduction and multivariate analysis [20]. PCA has wide applications in the research areas of data processing and analysis, pattern recognition, data visualization, image processing, etc.,

Step 1: Input Feature Data.

Step 2: Calculate mean of the data.

Let X= {x1, x2, x3,.....,xn}, Y= {y1, y2,y3,.....,yn}

$$mean = \frac{x1 + x2 + x3 + \ldots\ldots + xn}{n} \quad (1)$$

where n is total number of data points in dataset.

Step 3: Measure deviation

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x - \bar{x})^2} \quad (2)$$

Step 4: Calculate the Co-variance

$$\text{Cov}(X,Y) = \sum_{j=1}^{n} \frac{(xj - \mu x)(yj - \mu y)}{n-1} \quad (3)$$

Step 5: Calculate Eigen Values and Eigen Vectors.

Let A be an n n matrix. The eigenvalues of A are the roots of the characteristic polynomial

p( )=det(A− I)

For each eigen value , we find eigenvectors [v=] by solving the linear system

(A− I)v=0

The set of all vectors v satisfying Av= v is called the eigen space of A corresponding to .

Step 6: Principal Feature space

The importance of PCA comes from three main properties:

1. PCA is the most favorable linear scheme for compressing a group of high dimensional vectors into a group of lower dimensional vectors and then rebuilding the original set.
2. The model parameters can be computed directly from the data [20].
3. Compression and decompression are simple operations to carry out given the model parameters - they involve computing matrix multiplication.

PCA summarizes the variation in correlated multivariate attributes to a set of non-correlated components, each of which is a particular linear combination of the original variables. The extracted non-correlated components are called Principal Components (PC) and are estimated from the eigenvectors of the covariance matrix of the original variable [21]. Therefore, the objective of PCA is to achieve parsimony and reduce dimensionality by extracting the smallest number components that account for most of the variation in the original multivariate data and to summarize the data with little loss of information.

The model which uses PCA detects and identifies intrusions by profiling normal network behavior as well as various attack behaviors. This is very useful for preventing intrusions according to the associated individual type of attack. The model can also achieve real-time intrusion identification based on dimensionality reduction and on a simple classifier.

K- means clustering :

To create clusters from the input data, we have used K-means clustering algorithm. k- means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The algorithm initially have empty set of clusters and updates it as proceeds. For each record it computes the Euclidean distance between it and each of the centroids of the clusters. The instance is placed in the cluster from which it has shortest distance. Assume we have fixed metric M, and constant cluster Width W. Let dist(C, d) is the distance with metric M, Cluster centroid C and instance d where centroid of cluster is the instance from feature vector.

Algorithm 1: The K-means clustering algorithm

Let Y = {y1, y2, y3, ..., yi, ..., yn } // Set of n data points.

k // Number of desired clusters

Ensure: A set of k clusters.

Step 1: Arbitrarily choose k data points from Y as initial centroids;

Step 2: Repeat

Assign each point yi to the cluster which has the closest centroid;

Calculate the new mean for each cluster;

Until convergence criteria is met.

However, K-means algorithm has few limitations:

1. Initially K value assumption is very important but proper description is not provided when assuming k value.. Hence we get different number of clusters for different K values.
2. The initial cluster centroids are essential but if the centroid is far from the cluster center of the data itself then it results into infinite iterations which sometimes lead to incorrect clustering .
3. The K-means clustering is not good enough with clustering data set with noise [16].

PROPOSED METHODOLOGY

The proposed model consists of four phases: preprocessing, feature selection, classification, and evaluation of results which is shown in Figure 1. This model uses KDD dataset [22] for the experiment in this work. This dataset is a standard, which is considered as a benchmark for evaluating security detection mechanisms. It is a refined form of KDD cup dataset. The details of proposed model with each of its phases are described in the following subsections.

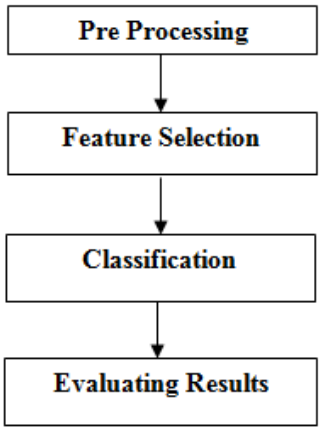

Figure 1: Block diagram

Pre-processing:

Preprocessing module helps to remove the redundant data, identify outliers, incomplete data and transforms the data into a meaningful format.

In preprocessing step we perform the following:

i. Check for null values and handle them (if any).

ii. Convert categorical data to numerical data.

Feature Selection: This module prepares the data for classification and removed irrelevant features and duplicate instances. We use PCA approach for retrieving the relevant features from the dataset. In the proposed method, each network connection is transformed into an input data vector. PCA is employed to reduce the high dimensional data vectors and identification is thus handled in a low dimensional space with high efficiency and low usage of system resources. The distance between a vector and its reconstruction onto those reduced subspaces representing different types of attacks and normal activities is used for identification. The little computational cost of the distance allows a real-time performance of intrusion detection.

Principal Component Analysis is a feature extraction technique that generates new features which are linear combination of the initial features. PCA maps each instance of the given dataset present in a d dimensional space to a k dimensional subspace such that k < d. [15] The new dimensions generated are called the Principal Components & each principal component is directed towards maximum variance apart from the variance already accounted for in all its earlier components. Consequently, the primary component covers the maximum variance and other components covers lesser value of variance.

Classification: This module performs classification using enhanced K-means algorithm to enhance the classification accuracy.

Algorithm 2: The Enhanced k- means Method

Input: Y = {y1, y2, y3, ..., yi, ..., yn } // set of n numbers of data points [16]

k // The number of desire Clusters

Output: A set of k clusters

Part1: Find out initial centroids

Step 1.1: Get the mean value for the given dataset.

Step 1.2: Find the distance for each data point from obtained mean value.

Step 1.3: Sort data points according to their distance from the mean value computed in step 2.1.

Step 1.4: Obtain K number of equal subsets from data set.

Step 1.5: Calculate the middle point for each subset which will become initial centroids.

Step 1.6: Compute distance from each data point to initial centroids [17].

REPEAT

Part 2: Assigning data points to nearest centroids

Step 2.1: Calculate distance from each data point to centroids and allot data points to its nearest centroid to form clusters.

Step 2.2: Calculate new centroids for the obtained clusters.

Step 2.3: Compute distance from all centroids to each data point for all the conditions.

IF

The Distance >= distance stored previously,

THEN

These data points doesn't needs to shift to other clusters.

ELSE

From the distance calculated assign data point to its nearest centroid by comparing distance from different centroids.

Step 2.4: Calculate centroids for these new clusters again Until the convergence criterion met. Final result: A Set of K clusters [18].

EVALUATING RESULTS

We evaluated this approach on intrusion KDDCup'99 benchmark dataset. In this paper, 10% KDD Cup'99 dataset [12] is used for experimentation.

Data Source and Dataset Description:

In this section, we provide a brief description of KDDCup'99 dataset [10] which is derived from UCI Machine Learning Repository [11] [22]. In 1998, DARPA intrusion detection evaluation program, to perform a comparison of various intrusion detection methods, a simulated environment, was set up by the MIT Lincoln Lab to obtain raw TCP/IP dump data for a local-area network (LAN). The functioning of the environment was like a real one, which included both background network traffic and wide variety of attacks.

In PCA it is very important to identify how many principal components (PCs) should be retained to account for most of the data variability i.e., the dimension of the subspace. In other words, the number of features should be considered to make the connections more discriminate. In this paper a 10 fold cross validation is applied on KDD dataset for the selection of the number of features.

Performance Parameters:

There are many measures available for evaluating system performance. For evaluating intrusion detection results following measure are generally used.

1. True positive (TP): TP is the number of normal i.e., genuine connections that were correctly classified as normal

among the total number of genuine connections.

$$TP = \frac{C1}{P} \qquad (4)$$

where C1 is the number of normal i.e., genuine connections that were correctly classified as normal and P is the total number of genuine connections tested.

2. False positive (FP): FP is the number of intrusion connections that were correctly classified as intruder among the total number of intrusion connections.

$$FP = \frac{C2}{N} \qquad (5)$$

where C2 is the number of intrusion connections that were correctly classified as intruder and N is the total number of intrusion connections tested.

3. Accuracy: Accuracy is the number of total connections that were correctly classified.

$$\text{Accuracy} = \frac{TP + FP}{2} \qquad (6)$$

Results :

The experimental results from this approach are listed in Table 1. By varying the number of PCA features, we observed TP, FP and accuracy. It is observed that, the performance increases with decreasing number of PCA features from 35 to 15. However, further decreasing the number of features lead to diminishing performance

Hence, in the experiments the optimal value for number of PCA features is chosen as 15. Further, we also compared the proposed system with the k - means clustering algorithm. This is shown in Table 2. It is observed that, the proposed approach performs well as shown in figure: 2(a), figure: 2(b), figure: 2(c), figure: 2(d).

Table 1: Performance analysis with varying number of features selected using PCA.

Number of features	TP	FP	Accuracy
41	0.9	0.92	0.91
35	0.91	0.936	0.923
25	0.915	0.941	0.928
15	0.926	0.938	0.932
10	0.913	0.935	0.924

Table 2: Performance comparison with K-means clustering approach.

Number of features	K-means clustering	Enhanced K-means clustering
41	0.89	0.91
35	0.904	0.923
25	0.919	0.928
15	0.923	0.932
10	0.918	0.924

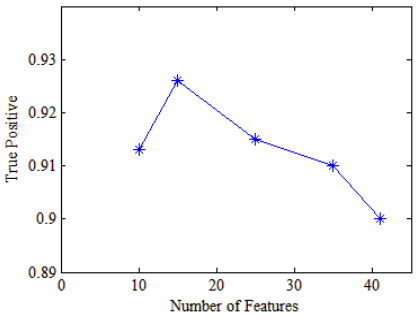

Figure 2: (a) True Positive

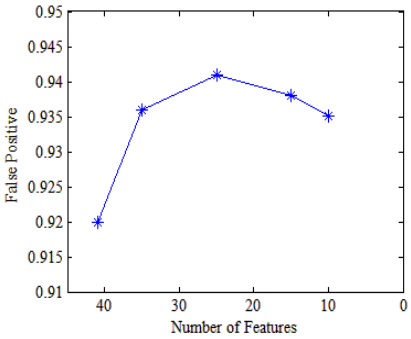

Figure 2: (b) False Positive

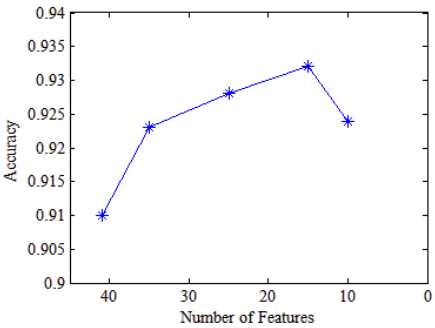

Figure 2: (c) Accuracy

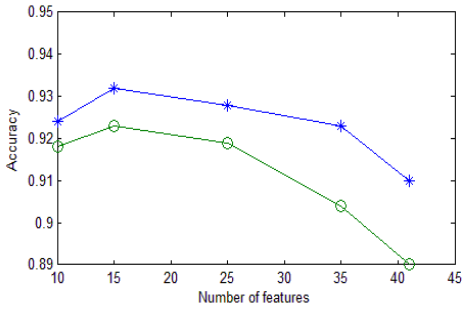

Figure: 2 (d) Accuracy comparison of K-means and enhanced k means

CONCLUSIONS

In recent years, many research has been done to develop an effective data mining-based IDS. An effective IDS is defined when it can simultaneously obtain both high classification accuracy. In this study, we investigated the effectiveness of enhanced K-means clustering algorithm for intrusion detection. Our experiment shows that along with PCA algorithm, the enhanced K-means algorithm performs better and achieves more than 90% accuracy in intrusion detection.

ACKNOWLEDGMENT

I heartily thank UCI Machine Learning repository for providing the KDD CUP 1999 dataset which help me to work on this research.

REFERENCES

[1] Kou, G et al, Multiple criteria mathematical programming for multi-class classification and application in network intrusion detection. Information Sciences 2009; 179(4): p.371-381.

[2] Tsai, C.-F., et al., Intrusion detection by machine learning: A review. Expert Systems with Applications 2009; 36(10): p. 11994-12000.

[3] Chung, YY. and Wahid N, A hybrid network intrusion detection system using simplified swarm optimization (SSO). Applied Soft Computing, 2012; 12(9): p. 3014-3022.

[4] Wu, SX. and Banzhaf W, The use of computational intelligence in intrusion detection systems: A review. Applied Soft Computing, 2010; 10(1): p. 1-35.

[5] Patel, A, et al, An intrusion detection and prevention system in cloud computing: A systematic review. Journal of Network and Computer Applications, 2013; 36(1): p. 25-41.

[6] Mohammed, MN and Sulaiman N, "Intrusion detection system based on SVM for WLAN" Procedia Technology, 2012;1: p. 313-317.

[7] Chen, WH, Hsu H.S, and Shen H.P, Application of SVM and ANN for intrusion detection. Computers & Operations Research, 2005; 32(10): p. 2617-2634.

[8] Wang, G, et al, "A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering". Expert Systems with Applications, 2010; 37(9): p. 6225-6232.

[9] Kolias, C, Kambourakis G, and Maragoudakis M, "Swarm intelligence in intrusion detection: A survey" computers & security, 2011; 30(8): p. 625-642.

[10] C. B. D. Newman and C. Merz, "UCI repository of machine learning databases," Tech. Rep.,Department of Information and Computer Science, University of California, Irvine, Calif, USA, 1998, http://www.ics.uci.edu/mlearn/MLRepository.

[11] M. Tavallaee, E. Bagheri,W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA '09), July 2009.

[12] P. Amudha, S. Karthik, and S. Sivakumari "A Hybrid Swarm Intelligence Algorithm for Intrusion Detection Using Significant Features" Hindawi Publishing Corporation e Scientific World Journal Volume 2015, Article ID 574589, 15 pages http://dx.doi.org/10.1155/2015/574589.

[13] Kamini Nalavade, B. B. Mehsram, Ph.D "Evaluation of K-means Clustering for Effective Intrusion Detection and Prevention in Massive Network Traffic Data" International Journal of Computer Applications (0975 – 8887) Volume 96– No.7, June 2014.

[14] Iftikhar Ahmad, "Feature Selection Using Particle Swarm Optimization in Intrusion Detection" Hindawi Publishing Corporation International Journal of Distributed Sensor Networks Volume 2015, Article ID 806954, 8 pages http://dx.doi.org/10.1155 /2015/806954.

[15] K. Keerthi Vasan , B. Surendiran " Dimensionality reduction using Principal Component Analysis for network intrusion detection" Perspectives in Science (2016) 8, 510—512 http://dx.doi.org/10.1016/j.pisc.2016.05.010.

[16] C. Zhang, and Z. Fang, "An improved K-means clustering algorithm", Journal of Information & Computational Science, vol. 10(1), pp.193-199, 2013.

[17] M. Yedla, S.R. Pathakota, and T.M. Srinivasa, "Enhancing K-means Clustering algorithm with Improved Initial Center", International Journal of

Computer Science and Information Technologies, vol.1 (2), pp.121-125, 2010.

[18] S. Na, G. Yong, and L. Xumin, Research on K-means clustering algorithms, IEEE Computer society, vol.74, pp.63-67, 2010.

[19] Sundus Juma, Zaiton Muda, M.A. Mohamed, Warusia Yassin, "Machine Learning Techniques For Intrusion Detection System: A Review" Journal of Theoretical and Applied Information Technology 28th February 2015. Vol.72 No.3.

[20] Vipin Das, Vijaya Pathak, Sattvik Sharma, Sreevathsa, MVVNS Srikanth, Gireesh Kumar T "System Based On Machine Learning Algorithms" International Journal of Computer Science & Information Technology (IJCSIT), Vol 2, No 6, December 2010. 21.

[21] Awari, Mahesh Babu. URBAN TRAFFIC COMPLEXITY AND SOLUTIONS. Lulu. com 2016.

[22] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[23] N. Chandra Sekhar Reddy, Dr. Purna Chandra rao, Dr. A. Govardhan, "An Intrusion Detection System for Secure Distributed Local Action Detection and Retransmission of Packets", International Journal of Soft Computing 12(6): 123-129, 2016, ISSN: 1816-9503, © Medwell Journals, 2016