ISSN 2395-1621

# Enhanced detection rate through PCA and radial SVM in Wireless Sensor Networks

[#1]Riyazahmed A Jamadar, [#2]Prof. Ms.Mousami S Vanjale

[#1]riyaz.jamadar@gmail.com

[#]1Research Scholar, AISSMS IOIT, Pune-01
[#2]Research Scholar, Bharati Vidyapeet, Pune-45
.

**ABSTRACT**

**Security in Wireless sensor network(WSN) is a paramount aspect, as the WSN is vulnerable to a wide range of attacks due to deployment in the hostile and unattended environment with constrained resources. Intrusion detection system is one of the major and efficient defensive methods against various attacks in WSN. Anomaly based detection has emerged as a popular technique which leverages the state of the art machine learning algorithms to enhance the detection rate and reduces the false positive rate. In this paper a Principal Component Analysis is used for feature extraction and dimensionality reduction, and Radial Support Vector Machine for classification. The implemented algorithms demonstrate increased detection rate, reduced false positive rates and increased true positive rates.**

ARTICLE INFO

**Article History**

Received : 16th July 2015

Received in revised form :

18th July 2015

Accepted : 21st July 2015

**Published online :**

**29th July 2015**

**Keywords— detection rate, principal component analysis, classification.**

## I. INTRODUCTION

The advent of cutting edge technologies like VLSI and Wireless Communications have made Wireless sensor networks (WSNs) to develop feasible and affordable systems for military, health care and agriculture. Basically WSNs employ battery as a primary power source and harvest power from the environment like solar panels as a secondary power supply. Normally the aim of intruder is to steal the information or to create disorder in the functioning of the WSN there by targeting draining out of these resources.

Security and Privacy are important challenges in all types of wired and wireless communications. These will become more serious in wireless sensor networks, as the structure of these networks make them attractive targets for intrusions and other attacks. These attacks have serious consequences if any breach of security, compromise of information, or disruption of correct application behavior on, applications such as Defense/military and Disasters. As WSNs are normally deployed in remote and unattended areas, they provide an easy for intrusions. WSNs are typically very resource-constrained and operate in harsh environments, which further facilitate compromise and make it often difficult to distinguish security breaches from node failures, varying link qualities, and other commonly found challenges WSNs.

To detect various threats, intrusion detection systems are needed. There are two types of intrusion detection systems: host-based and network-based. Host-based technology examines events like what files were accessed and what applications were executed. Network-based technology examines events as packets of information exchange between computers. One of the main problems for NIDSs is to build effective behavior models to distinguish normal behaviors from abnormal behaviors by observing data [1,2].

An anomaly detection system is one of the intrusion detection systems available to model the normal system/network behavior which is effective in identifying both common as well as uncommon attacks. It is built on a normal system that analyzes the network or program activity. There are a number of different architectures and methods used for anomaly detection. They are statistical approach,

 

clustering approach, centralized approach, artificial immune system, isolation table, machine learning approach and game-theory approach. Existing intrusion detections are not sufficient in detecting the novel attacks. Therefore, some anomaly detection approaches work on the available normal data and model them to identify the deviations. Machine learning deals with ability of a program to train and enhance the performance on a certain task [4].

The classification of Machine-learning based anomaly detection is as follows:

- Anomaly detection based on supervised-learning
- Anomaly detection based on unsupervised-learning
- Anomaly detection based on semi-supervised-learning

## II. RELATED WORK

Under supervised learning, Wang Hui, Zhang Guiling et.al[6], describe a detection model based on Super Vector Machine(SVM) that combines Principal Component Analysis (PCA) and Particle Swarm Optimization for anomaly detection. PCA primarily performs dimension reduction and PSO algorithm is used to optimize the factors in SVM. This has moderately good detection rate with high computational overheads.

Weiming Hu and Steve Maybank, have proposed a detection model based on AdaBoost algorithm decision stumps that are used as weak classifier , and rules are provided for both categorical and continuous features. The data are classified for training that contains both normal data set and attack data set. This method has very low detection rate of 90.88[8].

AmuthanPrabakar et.al [9], have proposed a work which uses K-Means and C4.5 algorithms for distinguishing the normal and anomalous activities in a computer network. Here the process of cascading the K-Means and C4.5 method includes two phases: i) Selection phase ii) Classification phase. In the selection phase, Euclidean distance is measured to identify the closest cluster and C4.5 decision tree is employed to handle the neighbor cluster, whereas in the classification phase, C4.5 decision tree are computed on the test instance to classify it as normal one or anomalous one.

Weiming Hu. et.,al [10], developed a detection model called Online AdaBoost-based Intrusion Detection approach which uses these algorithms, i) AdaBoost algorithm and decision stumps ii) Online AdaBoost classifier and online Gaussian Mixture Models (GMM). These are considered as weak classifiers. The local parametric models for intrusion detection are shared between the nodes of the network. Particle swarm optimization and support vector machine are used to cascade the local detection models into a global detection model. The authors of this work claim that it has highest detection rate with moderately good false positive rate.

Unsupervised Learning-based Anomaly Detection, A hyper spherical cluster based detection algorithm is proposed by SutharshanRajasegarar et.al [11], In this approach, each sensor node sends all its data to the gateway node and combines all data to form a combined dataset. The fixed-width clustering algorithms are used for anomaly detection. They randomly choose data points as centroid and Euclidean distance is measured between the centroid and next remaining data vector. If the distance to the closest centroid from a data is less than radius, then the data is added to that cluster, otherwise a new cluster is formed. In the distributed approach, the model is scattered to all sensor nodes. The local anomalies are detected and clusters are classified as normal or anomalous using K-NN classifier.

A new distributed online anomaly detection model is proposed by MuradA.Rassamet.al to measure the dissimilarities of sensor observations in WSN [12]. The candid-covariance free principal component analysis is utilized for data reduction in WSN. It includes two main stages. They are: i) Training stage ii) Detection stage. In the training stage, the observations are gathered at every sensor node to find the local normal model and sent it to the cluster head to create a global normal model. The detection threshold that represents the local normal model is chosen as the maximum and minimum range, whereas in the detection stage, every sensor observation is classified as normal or anomalous by analyzing the detection threshold from global normal model. This algorithm is based on the distance similarity to find the global anomalies in WSN.

MuradA.Rassam et.al have proposed another work which is a principal component classifier-based anomaly detection model is designed to detect anomalous sensor measurements to track dynamic changes [10]. The model has three phases: i) Training phase, ii) online detection phase and iii) update phase. In the training phase, the standard data analysis is collected at each sensor node to frame normal model. The detection phase compares each data with normal model framed in the training phase to classify the data as normal or anomalous. In the update phase, the normal model is retrained to produce a new normal reference model. Online Adaptive PCA classifier has highest detection rate and lowest false alarm rate, but the authors cite that, it has the limitation of consuming more energy, as it needs collect the information online and keep updating and retrain and add it to the reference model.
The performance comparison of unsupervised anomaly detection algorithms in terms of detection rate (DR) and false alarm rate (FAR) are shown in Table.2

Annie George[13], proposed a data mining and machine learning based detection system called ,Anomaly Detection based on Machine Learning: Dimensionality Reduction using PCA and Classification using SVM, which exploits PCA for dimension reduction and Linear SVM for classification. The limitation of this method is that the detection rate is substantially low with high false positive alarm rate, moreover it produces inaccurate results on addition of slight noise.

All the existing proposed work desperately addresses the detection rates and false alarm rate. None of the proposed work emphasize the effect of heavy computation based

 

algorithms on energy consumption and memory requirements. Thus, our work uses Principal Component Analysis for dimension reduction , which primarily reduces the no.of features. This greatly reduces the execution time and conserves the energy of the sensor network. Our proposed system also uses radial binary SVM which improves the margin of classification, yielding enhanced detection rate and minimizing false alarm rate.

## III.PROPOSED SYSTEM

In this work we have used the KDD99 dataset which is meant for intrusion detection based on data mining algorithms, and was established by the Third International Knowledge Discovery and Data Mining Tools Competition [15]. In the KDD99 data set, each data record corresponds to a set of derived features of a connection in the network data. Each connection is marked either as normal or as an attack, with exactly one specific attack type[13].

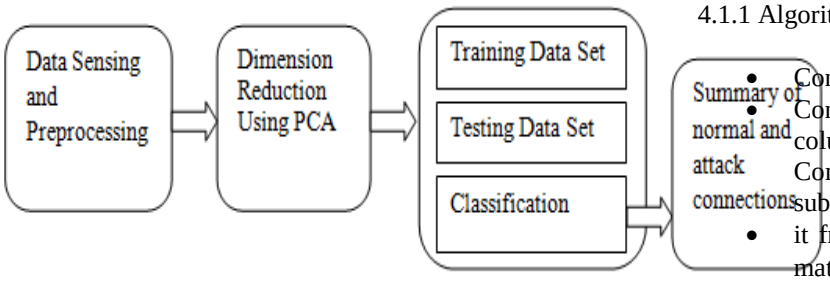


Fig1. Proposed IDS Framework

The technical challenges in Network Intrusion Detection based on machine learning methods are dimensionality reduction and classification. There are three main parts depicted in Proposed IDS Framework as shown in figure.1 for an intrusion or anomaly detection. Data Sensing and pre-processing of network data, feature extraction using PCA, and classification using radial SVM.

The benchmark KDD99 dataset consisting of connection records for each network connection represented by the forty two most important features derived from the network data [16]. From this labeled connection records we need to map the labels to numeric values so as to make it suitable to be the input of our machine learning algorithms. Also assign target class to the connections according to class label feature, which is the last feature in the connection record and assigns a target class zero for normal connection and a one for any deviation from that.

Dimension reduction is to be done for the given set of forty two features in the dataset. The PCA algorithm is considered [17]. Input is the set of connections represented by the forty two features. This module is important as we can represent the original features using a reduced feature set with maximum variance.

As the connection record in the KDD99 dataset consists of both nominal and binary data, a mapping mechanism is needed for the nominal data to make it suitable for the algorithms. Some of the nominal attributes in the record are the protocol name, the service and the class label. The mapping of these attributes is done using a dictionary with a unique value for each unique nominal attribute defined. While the nominal to numeric conversion is done, a comparison is done for each current attribute from the connection record and if a match is found from the dictionary, the nominal attribute is mapped to its corresponding value from the dictionary.

### 4.1 PRINCIPAL COMPONENT ANAYSIS FOR DIMENSION REDUCTION

PCA is one of the most fundamental tools of dimensionality reduction for extracting effective features from highdimensional vectors of input data [7, 8]. In this section, we will see the application of PCA for dimensionality reduction of network connection data consisting of forty two features, for making the classification problem more efficient.

#### 4.1.1 Algorithm

- Consider the network data corresponding to each
- Connection record after mapping. Thus each column represents a dimension of the input data. Compute the mean for each dimension, and subtract
- it from each data value. Compute the covariance matrix C of the input data• matrix. Calculate the Eigen values and the corresponding
- Eigen vectors for this covariance matrix, and the principal components are computed by solving the Eigen values problem of covariance matrix C. To find the principal components, choose the
- eigenvectors corresponding to K largest Eigen values, where K<<N

Dimensionality reduction step keep only the terms corresponding to the K largest Eigen values. Hence obtain a new feature vector consisting of eigenvectors of principal components. The final data computed using this feature vector and the mean adjusted original input data using the given equation

$$FinalData = RowFeatureVector \times RowDataAdjust$$

RowFeatureVector is the matrix in which eigenvectors in the columns transposed and RowDataAdjust is the mean adjusted input data. The obtained subspace is spanned by the orthogonal set of eigenvectors which reveal the maximum variance in the data space. Using PCA mapping high-dimensional data into low dimensional data reduces the calculation cost of NIDS and improves the efficiency of the analysis. Here principal component analysis has been used for dimensionality reduction of the forty two dimensions and the output of PCA method provides a set of features that are the linear combination of the original set of features. It accomplishes this by projecting data from a higher dimensional space to lower dimensional space such that error incurred by reconstructing the data in higher dimension is minimized. Thus the input for the SVM becomes more efficient as it represents the principal components that are with maximum variance and that are

 

orthogonal, thereby making the new sub space consisting of features somewhat clustered according to variance and hence the classification by discriminating plane which considers minimum variance becomes more accurate. When viewed from an informative view point, PCA provides SVM with the features that provide efficient classification.

## IV. RADIAL SVM FOR CLASSIFICAON

Use of radial SVM results in obtaining better results from the classification process when compared to normal linear SVM. In linear SVM, the classification is made by use of linear hyper-planes where as in radial SVM, nonlinear kernel functions are used and the resulting maximum-margin hyper-plane fits in a transformed feature space. The corresponding feature space is a Hilbert space of infinite dimensions, when the kernel used is a Gaussian radial basis function. The Gaussian Radial Basics function is given by the equation:

$\Phi(X\text{-}Xj) = \exp((-1/2\varphi^2_j) \|x-x_j\|^2)$ j=1,2, … N

Where j=1,2,….N. The $j^{th}$ input data point $x_j$ defines the center of radial basis function, the vector x is the pattern applied to the input. Φj is a measure of width of $j^{th}$ Gaussian function with center $x_j$.

We provide the SVM algorithm the input that includes the target class, and then the above steps are executed for the training dataset. It calculates the margin, the support vectors, the alpha values and then the weights. For our connection records, class labels as 0 for normal and 1 for anomaly class is assigned. This phase generates a training model for the data.

In the testing phase, we provide the test dataset without the target class. This phase considers the model generated by training for classification problem. For classification, the voting method is used, where for each input set, the class having maximum votes is considered. Then the input data belongs to that class. Here vote represents the decision of each binary classification. For our connection records, the classification will be as whether each record is normal or an anomaly.

## V. EXPERIMENTAL RESULT

Here in this work we perform the evaluation of our machine learning methods using KDD99 dataset. We are considering a set of connection records from the training and testing data of KDD99 dataset for evaluating our algorithms. We are comparing the performance based on SVM and SVM with PCA algorithms. When we consider the PCA for dimensionality reduction, the entire set of forty two features are considered as the input. The PCA finds the principal components among the set of features that are having the largest Eigen values. The PCA provides a linear combination of the original features as selected features which are uncorrelated with one another. The results show that the feature set is reduced to twenty eight retaining the class label feature. Even though the output is a linear combination of features, the class label feature doesn't change. The low dimensional data corresponding to the maximum variance principal components are used for classification.

Table 1. Comparison of various anomaly based detection with our system

| S.No | IDS algorithms | Detection rate | False alarm rate |
|---|---|---|---|
| 1 | PCA, SVM and PSO | 97.75 | Not mentioned |
| 2 | AdaBoost algorithm | 90.88 | 1.7 |
| 3 | K-Means and C4.5 algorithms | 99.6 | 0.1 |
| 4 | Online AdaBoost-based | 99.99 | 0.39 |
| 5 | A hyper spherical cluster | 85.47 | 1.48 |
| 6 | PCA based | 82.86 | 13.3 |
| 7 | Online Adaptive PCA classifier | 97.84 | 1.10 |
| 8 | Our Proposed Work PCA with radial SVM | 97.85 | 0.54 |

As the components are orthogonal in the new subspace where the features are grouped according to maximum variance, it enhances the classification algorithm to find a linear discriminating plane by classifying the features according to minimum variance, thereby increasing accuracy by decreasing the number of misclassification. The classification algorithm identifies any anomaly in the test data according to the training model. SVM is evaluated using the original data and the reduced data for performance analysis.

Using Principal Component Analysis for the data after pre-processing increases the classification accuracy of the network data as it finds the principal components which improves the linear separability of the data. Evaluation based on the SVM with PCA approach gives less misclassification compared to SVM method.

The SVM method uses the original set of forty two features and SVM with PCA uses the mapped set of twenty eight features with class label retained. Performances of both algorithms for anomaly detection are evaluated based on the precision and recall parameters. But the execution speed of second method is more as it takes less time with reduced feature set in the orthogonal space where the classification of the features is enhanced by the maximum variance components that are grouped thereby increasing classification speed, and the result is shown in Table 3. Comparison of both the methods shows that using PCA for reducing the high dimensional network data improves the speed of the detection system by enhancing the feature classification which is very important for an intrusion detection system.

Table 2 shows the performance of the classification methods based on precision and recall for anomaly and normal classes. Figure 2 shows the comparison results for both

 

categories and we can observe that latter one shows higher precision and recall value which explains that it causes only one or two misclassifications as the values depends on the correctly classified samples and hence more accurate. The parameter values increases according to the number of correctly classified samples which shows its accuracy.

Table 1.Performance based on Table 2. Performance based on precision and recall
execution time.

| Classifiers | Precision | Recall |
|---|---|---|
| SVM | 0.81 | 0.802 |
| SVM with PCA | 0.97 | 0.9 |

| Classifiers | Execution TIme |
|---|---|
| SVM (42 features) | 0.291 |
| Radial SVM with PCA (28 features) | 0.0088 |

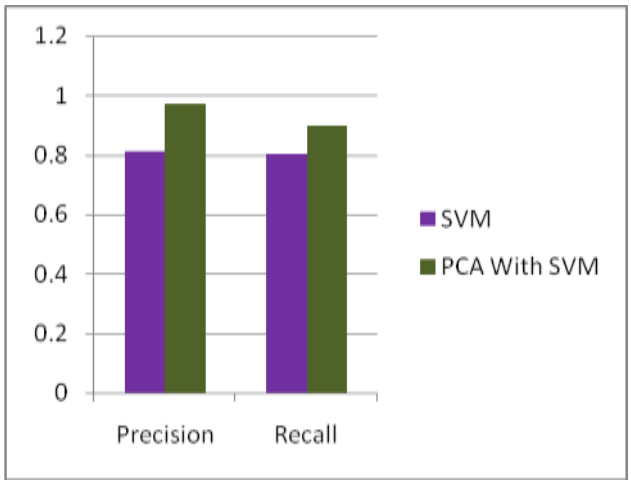


Fig2. Comparison of classification methods using Precision and Recall

As shown in Fig 2. the comparison of results both categories, we can observe that latter one shows higher precision and recall value which explains that it causes only one or two misclassifications as the values depends on the correctly classified samples and hence more accurate. The parameter values increases according to the number of correctly classified samples which shows its accuracy.

## VI. CONCLUSION

The work examines an anomaly detection system using machine learning algorithms such as Principal Component Analysis and Support Vector Machine. The KDD99 connection record has been converted into the required format for the machine learning algorithms using a mapping technique that is important in case of any machine learning algorithm. Dimensionality reduction with PCA helps to reduce high dimensional network data to provide the more informative features from the data thereby decreasing the execution time for classification and also increasing the classification accuracy.

Support Vector Machine (SVM) helps to classify our reduced network data to detect it as a normal or an anomaly connection. The generalization concept can help to obtain better classification result. Using PCA with SVM provides higher accuracy as the output of PCA is the maximum variance components which can be efficiently separated using the hyper plane. We can also consider multiple classes in the anomaly class as we use multi-class SVM for classification.

The experimental results have been analyzed based on precision and recall values for each class and it shows that classification using dimensionality reduction is more accurate depending on the new subspace where the features are combined together according to maximum variance which enhances classification using discriminating plane. Further work can be done for more types of anomalies that are emerging at present, which helps the NIDS to be more efficient. The algorithms can be used for anomaly detection in other application areas also as here it is for network data.

## REFRENCES

[1]. W.K. Lee, S.J.Stolfo, “A data mining framework for building intrusion detection model”, *In: Gong L., Reiter M.K. (eds.): Proceedings of the IEEE Symposium on Security and Privacy*. Oakland, CA: IEEE Computer Society Press, 1999.

[2]. W.K. Lee, et al., "Mining audit data to build intrusion detection models", In Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'98), pp.66-72, 1998 .

[3].H. GüneşKayacık, A.NurZincir-Heywood, Malcolm I. Heywood, “Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets”, Dalhousie University, Faculty of Computer Science, 6050 University Avenue, Halifax, Nova Scotia.

[4]. AnimeshPatcha and Jung-Min Park, “An overview of anomaly detection techniques: Existing solutions and latest technological trends”, *Elsevier Computer Networks, Vol. 51,* 2007.

[5] MahbodTavallaee,EbrahimBagheri, Wei Lu, and Ali A. Ghorbani, “A Detailed Analysis of the KDD CUP 99 Data Set “, *Proceedings of the IEEE Symposium on Computational Intelligence in Security and Security Applications*, 2009.

[6] Wang Hui, Zhang Guiling et.al, “A Novel Intrusion Detection Method Based on Improved SVM by Combining PCA and PSO”, Journal of Natural Sciences, Vol.16, No.5, 2011.

[7].Weiming Hu and Steve Maybank, “AdaBoost-Based Algorithm for Network Intrusion Detection”, IEEE, Vol.8, No.32, 2008.

 

[8]. J. Saranya, Dr.G.Padmavathi, "A Brief Study on Different Intrusions and Machine Learning-based Anomaly Detection Methods in Wireless Sensor Networks", Int. J. Advanced Networking and Applications Volume: 6 Issue: 4 Pages: 2414-2421 (2015) ISSN: 0975-0290

[9]. AmuthanPrabakarMuniyandi, R. Rajeshwari and R.Rajaram, Network Anomaly Detection by CascadingK-Means Clustering and C4.5 Decision Tree algorithm, *Elsevier, Vol.30*, 2012.

[10].Weiming Hu, Jun Gao, Yanguo Wang, Ou Wu and Steve Maybank, Online AdaBoost-Based Parameterized Methods for Dynamic Distributed Network Intrusion Detection, *IEEE, Vol.44, No.1*, 2014.

[11].SutharshanRajasegarar, Christopher Leckie and MarimuthuPalaniswami, "Hyper spherical cluster based distributed anomaly detection in wireless sensor networks", *Elsevier Journal of Distributed Computing,Vol.74*, 2014.

[12]. MuradA.Rassam, AnazidaZainal and MohdAizainiMaarof, An Efficient distributed anomaly detection model for wireless sensor networks, Elsevier, 2013.

[13]. Annie George, "Anomaly Detection based on Machine Learning: Dimensionality Reduction using PCA and Classification using SVM", International Journal of Computer Applications (0975 – 8887) Volume 47– No.21, June 2012.

[14]. A.M.Chandrasekhar and K.Raghuveer, "Intrusion Detection Technique by using K-means, Fuzzy Neural Network and SVM classifiers", presented at International Conference on Computer Communication and Informatics (ICCCI-2013), Coimbatore, INDIA.

[15]. H. GüneşKayacık, A.NurZincir-Heywood, Malcolm I. Heywood, "Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets", Dalhousie University, Faculty of Computer Science, 6050 University Avenue, Halifax, Nova Scotia

[16]. MahbodTavallaee,EbrahimBagheri, Wei Lu, and Ali A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set ", Proceedings of the IEEE Symposium on Computational Intelligence in Security and Security Applications, 2009.

[17]. Fengxi Song, ZhongweiGuo, Dayong Mei, "Feature selection using principal component analysis", Department of Automation and Simulation New Star Research Inst. Of Applied Tech. in Hefei City Hefei, China, International Conference on System Science, Engineering Design and Manufacturing Informatization, 2010.

 