Chapter 9
Applying Weighted PCA on Multiclass Classification for Intrusion Detection

Mohsen Moshki
Iran University of Science and Technology, Iran

Mehran Garmehi
Iran University of Science and Technology, Iran

Peyman Kabiri
Iran University of Science and Technology, Iran

ABSTRACT

In this chapter, application of Principal Component Analysis (PCA) and one of its extensions on intrusion detection is investigated. This extended version of PCA is modified to cover an important shortcoming of traditional PCA. In order to evaluate these modifications, it is mathematically proved that these modifications are beneficial and later on a known dataset such as the DARPA99 dataset is used to verify results experimentally. To verify this approach, initially the traditional PCA is used to preprocess the dataset. Later on, using a simple classifier such as KNN, the effectiveness of the multiclass classification is studied. In the reported work, instead of traditional PCA, a revised version of PCA named Weighted PCA (WPCA) will be used for feature extraction. The results from applying the aforementioned method to the DARPA99 dataset show that this approach results in better accuracy than the traditional PCA when a number of features are limited, a number of classes are large, and a population of classes is unbalanced. In some situations WPCA outperforms traditional PCA by more than 1% in accuracy.

DOI: 10.4018/978-1-60960-836-1.ch009
INTRODUCTION

Feature selection and feature reduction techniques are two major approaches to increase the performance of pattern recognition systems. Processing power needed by supervised and unsupervised learning algorithms has a close and direct relation with the number of features within the dataset. If it was possible to eliminate some of the less informative features or to reduce the total number of features, then it would lead to a dramatic reduction in the processing power required by pattern recognition systems used for intrusion detection. The advantages of feature reduction are not limited to faster processing but in some cases it may result in more accuracy. When the size of dataset is small, large number of the features may confuse the classifier. Some classifiers such as Multi Layer Perceptron (MLP) are sensitive to the ratio of the number of features to the number of database records. In these cases, feature reduction can be used as a solution to decrease this ratio for better classification.

Principal Component Analysis (PCA) is a well known feature extraction and reduction algorithm. PCA is an unsupervised algorithm and its transformation process is neither related to the population of the dataset nor number of the classes in the dataset. Although PCA extracts new features that make it possible to discriminate patterns more precisely, but there is no guarantee to increase class discrepancy. In this work, a simple modification to traditional PCA is used to remove this shortcoming. This version of PCA is a supervised extension that uses a weighting scheme by considering population of classes. After applying PCA to reduce the number of features, K-Nearest Neighbor (KNN) is used to classify samples. KNN is a simple instance-based method for object classification. Classification is a supervised task which tags new instances according to previously instances encountered by a classifier. KNN classifies each instance based on the learned pattern from its neighbors. In fact, this approach classifies each input based on the votes given by its neighbors.

First section of this paper is devoted to the related works in this area of research. After getting familiar with the scope of the problem, traditional PCA as a feature reduction method and KNN as a general purpose classifier will be introduced. In the following sections a simple and essential modification on PCA will be introduced. It will be proved that this modification is beneficial, and to do so, some practical studies will be presented. Finally, chapter will be concluded by presenting results and drawing the path to the future studies.

Related Works

One of the most challenging issues in intrusion detection and response is the necessary processing power. Most of the time the performance and accuracy are two sides of a coin. As a result of this fact the network security research communities
A Mark-Up Language for the Specification of Information Security Governance Requirements
www.igi-global.com/article/mark-language-specification-information-security/55378?camid=4v1a