Chapter 2

Improving Effectiveness of Intrusion Detection by Correlation Feature Selection

Hai Thanh Nguyen
Gjøvik University College, Norway

Katrin Franke
Gjøvik University College, Norway

Slobodan Petrović
Gjøvik University College, Norway

ABSTRACT

In this paper, the authors propose a new feature selection procedure for intrusion detection, which is based on filter method used in machine learning. They focus on Correlation Feature Selection (CFS) and transform the problem of feature selection by means of CFS measure into a mixed 0–1 linear programming problem with a number of constraints and variables that is linear in the number of full set features. The mixed 0–1 linear programming problem can then be solved by using branch-and-bound algorithm. This feature selection algorithm was compared experimentally with the best-first-CFS and the genetic-algorithm-CFS methods regarding the feature selection capabilities. Classification accuracies obtained after the feature selection by means of the C4.5 and the BayesNet over the KDD CUP’99 dataset were also tested. Experiments show that the authors’ method outperforms the best-first-CFS and the genetic-algorithm-CFS methods by removing much more redundant features while keeping the classification accuracies or getting better performances.

DOI: 10.4018/978-1-4666-2163-3.ch002
INTRODUCTION

Intrusion detection systems (IDS) have become important security tools applied in many contemporary network environments. They gather and analyze information from various sources on hosts and networks in order to identify suspicious activities and generate alerts for an operator. The task of intrusion detection is often analyzed as a pattern recognition problem - an IDS has to tell normal from abnormal behaviour. It is also of interest to further classify abnormal behaviour in order to undertake adequate counter-measures. An IDS can be modeled in various ways (Crescenzo et al., 2005; Gu et al., 2006). A model of this kind usually includes the representation algorithm (for representing incoming data in the space of selected features) and the classification algorithm (for mapping the feature vector representation of the incoming data to elements of a certain set of values, e.g., normal or abnormal etc.). Some IDS, like the models presented in (Gu et al., 2006), also include the feature selection algorithm, which determines the features to be used by the representation algorithm. Even if the feature selection algorithm is not included in the model directly, it is always assumed that such an algorithm is run before the very intrusion detection process.

The quality of the feature selection algorithm is one of the most important factors that affect the effectiveness of an IDS. The goal of the algorithm is to determine the most relevant features of the incoming traffic, whose monitoring would ensure reliable detection of abnormal behaviour. Since the effectiveness of the classification algorithm heavily depends on the number of features, it is of interest to minimize the cardinality of the set of selected features, without dropping potential indicators of abnormal behaviour. Obviously, determining a good set of features is not an easy task. The most of the work in practice is still done manually and the feature selection algorithm depends too much on expert knowledge. Automatic feature selection for intrusion detection remains therefore a great research challenge.

In this paper, we propose an automatic feature selection procedure for intrusion detection purposes based on so-called filter method (Guyon et al., 2006; Liu & Motoda, 2008) used in machine learning. The filter method directly considers statistical characteristics of the data set, such as correlation between a feature and a class or inter-correlation between features, without involving any learning algorithm. We focus on one of the most important filter methods, the Correlation Feature Selection (CFS) measure proposed by Hall (1999).

The CFS measure considers correlation between a feature and a class and inter-correlation between features at the same time. This measure is used successfully in test theory (Ghiselli, 1964) for predicting an external variable of interest. In feature selection, the CFS measure is combined with some search strategies, such as brute force, best first search or genetic algorithm, in order to find the most relevant subset of features. However, the brute force method can only be applied when the number of features is small. When the number of features is large, this method requires huge computational resources. For example, with 50 features the brute force method needs to scan all \(2^{50}\) possible subsets of features. That is impractical in general. With best first search or genetic algorithm, we can deal with high dimensional data sets, but these methods usually give us locally optimal solutions. It is desirable to get globally optimal subset of relevant features by means of the CFS measure with the hope of removing more redundant features and still keeping classification accuracies or even getting better performances.

The feature selection method that we propose in this paper finds the globally optimal subset of relevant features by means of the CFS measure. Firstly, we formulate the problem of feature selection by representing the CFS measure as an