Chapter XLIII
Identification Through Data Mining

Diego Liberati
Italian National Research Council, Italy

ABSTRACT

Four main general purpose approaches inferring knowledge from data are presented as a useful pool of at least partially complementary techniques also in the cyber intrusion identification context. In order to reduce the dimensionality of the problem, the most salient variables can be selected by cascading to a K-means a Divisive Partitioning of data orthogonal to the Principal Directions. A rule induction method based on logical circuits synthesis after proper binarization of the original variables proves to be also able to further prune redundant variables, besides identifying logical relationships among them in an understandable "if .. then .." form. Adaptive Bayesian networks are used to build a decision tree over the hierarchy of variables ordered by Minimum Description Length. Finally, Piece-Wise Affine Identification also provides a model of the dynamics of the process underlying the data, by detecting possible switches and changes of trends on the time course of the monitoring.

INTRODUCTION

In trying to detect cyber intrusions, it often turns out that one has to face a huge amount of data, which is often not completely homogeneous, and often without an immediate grasp of an underlying simple structure. Many records (i.e., logs from both authorized users and possible intruders) each instantiating many variables (like time, duration, Internet protocol (IP) address, and so on) are usually collected with the help of tracing tools.

Given the opportunity to have many logs on several possible intruders, one of the typical goals one has in mind is to classify subjects on the basis of a hopefully reduced meaningful subset of the measured variables.

The complexity of the problem makes it worthwhile to use automatic classification procedures.

Then, the question arises of reconstructing a synthetic mathematical model, capturing the most important relationships among variables, in order to both discriminate intruders from allowed users and possibly also infer rules of behavior that could help in identifying habits of some classes of intruders.

Such interrelated aspects will be the focus of the present contribution.

Four main general purpose approaches, also useful in the cyber intrusion identification context, will be briefly discussed in the present chapter as well as the underlying cost effectiveness of each one.

In order to reduce the dimensionality of the problem, thus simplifying both the computation and the subsequent understanding of the solution, the critical problems of selecting the most salient variables must be solved.

A very simple approach is to resort to cascading a divisive partitioning of data orthogonal to the principal directions divisive partitioning (PDDP) (Boley, 1998) already proven to be successful in the same context of analyzing the logs of an important telecommunications provider (Garatti, Savaresi, & Bittanti, 2004)

A possible approach that is more sophisticated is to resort to a rule induction method, like the one described in Muselli and Liberati (2000). Such a strategy also offers the advantage of extracting the underlying rules, implying conjunctions and/or disjunctions between the identified salient variables. Thus, a first guess of their even nonlinear relations is provided as a first step in designing a representative model, whose variables will be the selected ones. Such an approach has been shown (Muselli & Liberati, 2002) to be not less powerful over several benchmarks, than the popular decision tree developed by Quinlan (1994).

An alternative in this sense can be represented by adaptive Bayesian networks (Yarmus, 2003), whose advantage is that it is also available on a widespread commercial database tool like Oracle.

A possible approach to blindly build a simple linear approximating model is to resort to piece-wise affine (PWA) identification (Ferrari-Trecate, Muselli, Liberati, & Morari, 2003).

The joint use of (some of) these four approaches is described briefly in the present contribution, starting from data without known priors about their relationships, thus will allow reduction in dimensionality without significant loss in information, then to infer logical relationships, and, finally, to identify a simple input-output model of the involved process that also could be used for controlling purposes even in a critical field like cyber warfare.

BACKGROUND

The introduced tasks of selecting salient variables, identifying their relationships from data, and classifying possible intruders may be sequentially accomplished with various degrees of success in a variety of ways.

Principal components order the variables from the most salient to the least, but only under a linear framework.

Partial least squares do allow nonlinear models, provided that one has prior information on the structure of the involved nonlinearity; in fact, the regression equation needs to be written before identifying its parameters.

Clustering may operate even in an unsupervised way without the a priori correct classification of a training set (Boley, 1998).

Neural networks are known to learn the embedded rules with the indirect possibility (Taha & Ghosh, 1999) to make rules explicit or to underline the salient variables.

Decision trees (Quinlan, 1994) are a popular framework providing a satisfactory answer to the recalled needs.

MAIN THRUST OF THE CHAPTER

Unsupervised Clustering

In this chapter, we will firstly resort to a quite recently developed unsupervised clustering approach, the PDDP algorithm as proposed by Boley (1998).