CLASSIFYING DATA: THE GOOD, THE BAD, AND THE UGLY

A. BIEM, PhD
Vice President of Analytics,
Opera Systems

Opera Systems, NY, USA

Abstract. Data classification, namely the process of assigning a label or category to a data instance, is a one of a pre-requisites and pervasive steps in any Big Data application.

Application of data classification techniques can be found in most if not all data science applications. Customer analytics including customer targeting or customer retention models are exclusively based on method to classify data. In the medical domain, classification techniques permeate approaches for medical disease diagnosis, x-ray image analysis, or patient monitoring. Classification techniques are used every day for event detection in a variety of domain such as in cybersecurity, network analysis, or financial fraud detection. Document and text classification is the underlying basis for Natural Language Processing or social network analysis. We will provide a short survey of techniques used in data classification, assessing their strength and weaknesses and discussing their best use. The survey will cover several approaches to data classification including probabilistic classification, metric-based classification, rule-based classification, and techniques used for rare class learning.