ABSTRACT

SVDD Variants for Anomaly Detection with Implementations using Hadoop & Spark

Rekha A G, FPM 05/11/IT

Big data analytics facilitates better informed business decisions through the analysis of large data sets that remain unexploited by traditional business intelligence systems. ‘Big Data’ as input enhances the inferential power of established algorithms, but it challenges even the state-of-the-art computation and analysis methods. Though machine learning is a solution to overcome these problems, its current techniques have to be improved to deal with the Big Data. Another drawback of big data analytics is the greater focus on aggregates over outliers. However, in many situations the insights gathered from outliers could be of more significance. In light of this, the focus of this work is on developing machine learning techniques to make outlier detection practical on large business datasets. For over a decade, Support Vector Data Description (SVDD) technique has shown good predictive accuracy on a wide range of outlier detection tasks. It has been adapted to numerous business problems also. Inspired by this trend, this thesis explores the scalability problems associated with SVDD and tries to address it. Three approaches, namely, LT-SVDD, ELT-SVDD, and PELT- SVDD have been proposed. The feasibility of these methods was assessed using a set of experiments on synthetic as well as benchmark data sets; many of these with an order-of- magnitude advantage in terms of running time. The application of these methods to three real world business problems is also demonstrated. This work contributes to the support vector literature by establishing these methods as efficient for outlier detection on large data sets.