Mach Learn
DOI 10.1007/s10994-013-5422-z

An instance level analysis of data complexity

Michael R. Smith · Tony Martinez · Christophe Giraud-Carrier

Received: 6 January 2012 / Accepted: 26 September 2013

Abstract Most data complexity studies have focused on characterizing the complexity of the entire data set and do not provide information about individual instances. Knowing which instances are misclassified and understanding why they are misclassified and how they contribute to data set complexity can improve the learning process and could guide the future development of learning algorithms and data analysis methods. The goal of this paper is to better understand the data used in machine learning problems by identifying and analyzing the instances that are frequently misclassified by learning algorithms that have shown utility to date and are commonly used in practice. We identify instances that are hard to classify correctly (*instance hardness*) by classifying over 190,000 instances from 64 data sets with 9 learning algorithms. We then use a set of hardness measures to understand why some instances are harder to classify correctly than others. We find that class overlap is a principal contributor to instance hardness. We seek to integrate this information into the training process to alleviate the effects of class overlap and present ways that instance hardness can be used to improve learning.

Keywords Instance hardness · Dataset hardness · Data complexity

1 Introduction

It is widely acknowledged in machine learning that the performance of a learning algorithm is dependent on both its parameters and the training data. Yet, the bulk of algorithmic development has focused on adjusting model parameters without fully understanding the data that the learning algorithm is modeling. As such, algorithmic development for classification problems has largely been measured by classification accuracy, precision, or a similar metric on benchmark data sets. These metrics, however, only provide aggregate information about

Editor: Kiri Wagstaff.

M.R. Smith (✉) · T. Martinez · C. Giraud-Carrier
Department of Computer Science, Brigham Young University, Provo, UT, 84602, USA
e-mail: msmith@axon.cs.byu.edu

Published online: 05 November 2013