Mach Learn
DOI 10.1007/s10994-013-5412-1

Learning from multiple annotators with varying expertise

Yan Yan · Rómer Rosales · Glenn Fung · Ramanathan Subramanian · Jennifer Dy

Received: 2 August 2012 / Accepted: 19 August 2013

Abstract Learning from multiple annotators or knowledge sources has become an important problem in machine learning and data mining. This is in part due to the ease with which data can now be shared/collected among entities sharing a common goal, task, or data source; and additionally the need to aggregate and make inferences about the collected information. This paper focuses on the development of probabilistic approaches for statistical learning in this setting. It specially considers the case when annotators may be unreliable, but also when their expertise vary depending on the data they observe. That is, annotators may have better knowledge about different parts of the input space and therefore be inconsistently accurate across the task domain. The models developed address both the supervised and the semi-supervised settings and produce classification and annotator models that allow us to provide estimates of the true labels and annotator expertise when no ground-truth is available. In addition, we provide an analysis of the proposed models, tasks, and related practical problems under various scenarios. In particular, we address how to evaluate an-

Editors: Winter Mason, Jennifer Wortman Vaughan, and Hanna Wallach.

Y. Yan
Yahoo! Labs, Sunnyvale, CA 94085, USA
e-mail: chrisyan@yahoo-inc.com

R. Rosales (✉)
LinkedIn, Moutain View, CA 94043, USA
e-mail: rrosales@linkedin.com

G. Fung
Siemens Healthcare, Malvern, PA 19335, USA
e-mail: gfung@cs.wisc.edu

R. Subramanian · J. Dy
Electrical and Computer Engineering Department, Northeastern University, Boston, MA 02115, USA

R. Subramanian
e-mail: subramanian.r@husky.neu.edu

J. Dy
e-mail: jdy@ece.neu.edu

Published online: 19 October 2013