Mach Learn
DOI 10.1007/s10994-013-5411-2

Sequence labeling with multiple annotators

Filipe Rodrigues · Francisco Pereira · Bernardete Ribeiro

Received: 13 November 2012 / Accepted: 17 August 2013

Abstract The increasingly popular use of *Crowdsourcing* as a resource to obtain labeled data has been contributing to the wide awareness of the machine learning community to the problem of supervised learning from multiple annotators. Several approaches have been proposed to deal with this issue, but they disregard sequence labeling problems. However, these are very common, for example, among the Natural Language Processing and Bioinformatics communities. In this paper, we present a probabilistic approach for sequence labeling using Conditional Random Fields (CRF) for situations where label sequences from multiple annotators are available but there is no actual ground truth. The approach uses the Expectation-Maximization algorithm to jointly learn the CRF model parameters, the reliability of the annotators and the estimated ground truth. When it comes to performance, the proposed method (CRF-MA) significantly outperforms typical approaches such as majority voting.

Keywords Multiple annotators · Crowdsourcing · Conditional random fields · Latent variable models · Expectation maximization

Editor: Hal Daume, III.

F. Rodrigues (✉) · B. Ribeiro
Centre for Informatics and Systems of the University of Coimbra (CISUC), Department of Informatics Engineering, University of Coimbra, 3030-290 Coimbra, Portugal
e-mail: fmpr@dei.uc.pt

B. Ribeiro
e-mail: bribeiro@dei.uc.pt

F. Pereira
Singapore-MIT Alliance for Research and Technology (SMART), 1 CREATE Way, Singapore 138602, Singapore
e-mail: camara@smart.mit.edu

Published online: 04 October 2013