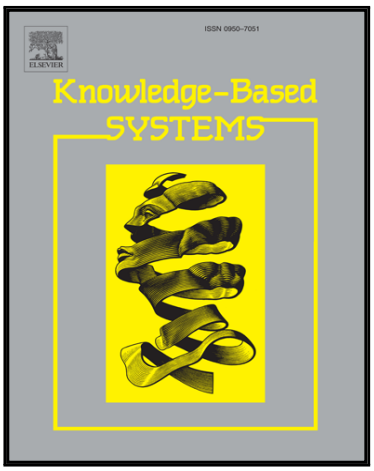
Accepted Manuscript



Active Learning with Confidence-based Answers for Crowdsourcing Labeling Tasks

Jinhua Song, Hao Wang, Yang Gao, Bo An

PII: S0950-7051(18)30346-0
DOI: 10.1016/j.knosys.2018.07.010
Reference: KNOSYS 4400

To appear in: *Knowledge-Based Systems*

Received date: 7 January 2018
Revised date: 21 June 2018
Accepted date: 2 July 2018

Please cite this article as: Jinhua Song, Hao Wang, Yang Gao, Bo An, Active Learning with Confidence-based Answers for Crowdsourcing Labeling Tasks, *Knowledge-Based Systems* (2018), doi: 10.1016/j.knosys.2018.07.010

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Active Learning with Confidence-based Answers for Crowdsourcing Labeling Tasks

Jinhua Song[a], Hao Wang[a], Yang Gao[a,*], Bo An[b]

[a]*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210093, China*
[b]*School of Computer Engineering, Nanyang Technological University, Blk N4-02c-110, Nanyang Avenue, 639798, Singapore*

Abstract

Collecting labels for data is important for many practical applications (e.g., data mining). However, this process can be expensive and time-consuming since it needs extensive efforts of domain experts. To decrease the cost, many recent works combine crowdsourcing, which outsources labeling tasks (usually in the form of questions) to a large group of non-expert workers, and active learning, which actively selects the best instances to be labeled, to acquire labeled datasets. However, for difficult tasks where workers are uncertain about their answers, asking for discrete labels might lead to poor performance due to the low-quality labels. In this paper, we design questions to get continuous worker responses which are more informative and contain workers' labels as well as their confidence. As crowd workers may make mistakes, multiple workers are hired to answer each question. Then, we propose a new aggregation method to integrate the responses. By considering workers' confidence information, the accuracy of integrated labels is improved. Furthermore, based on the new answers, we propose a novel active learning framework to iteratively select instances for "labeling". We define a score function for instance selection by combining the uncertainty derived from the classifier model and the uncertainty derived from the answer sets. The uncertainty derived from uncertain answers is more effective than that derived from labels. We also propose batch methods which select multiple instances at a time to further improve

*Corresponding author
Email addresses: songjinhua2008@gmail.com (Jinhua Song), wanghao.hku@gmail.com (Hao Wang), gaoy@nju.edu.cn (Yang Gao), boan@ntu.edu.sg (Bo An)

Download English Version:

https://daneshyari.com/en/article/10151093

Download Persian Version:

https://daneshyari.com/article/10151093

Daneshyari.com