論文 / 著書情報
Article / Book Information

題目(和文)	群衆の叡智からの統計的学習
Title(English)	Statistical Learning from the Wisdom of Crowds
著者(和文)	ZHANGHAO
Author(English)	Hao Zhang
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第10239号, 授与年月日:2016年3月26日, 学位の種別:課程博士, 審査員:杉山 将,徳永 健伸,篠田 浩一,村田 剛志,石田 貴士
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第10239号, Conferred date:2016/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

専攻： Department of	計算工学　　専攻	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of　（　工学　）
学生氏名： Student's Name	ZHANG HAO	指導教員（主）： Academic Advisor(main)	杉山 将
		指導教員（副）： Academic Advisor(sub)	藤井 敦

要旨（英文 800 語程度）
Thesis Summary（approx.800 English Words ）

Labeled data is essential to many machine learning applications. In recent years, crowdsourcing has emerged as a useful tool for combining human intelligence by asking a crowd of low-paid workers to complete a group of labeling tasks. This doctoral thesis is devoted to developing statistical methods of learning from the wisdom of crowds.

Since workers in crowds are usually non-experts, the collected labels often contain a significant amount of noises. To cope with this issue, many crowdsourcing methods jointly estimate the true labels and workers' reliability on the collected labels. However, they do not focus on how to collect these labels for the purpose of wisely using the total budget of label collection. This motivates us to consider another important problem called task-worker assignment, which is to repeatedly make decisions on which task is to assign to which worker. The task-worker assignment problem naturally includes two aspects: worker selection and task selection.

Worker selection aims to select reliable workers given any task. A challenge in this scenario is how to deal with the diversity of tasks and the involved trade-off between exploration and exploitation. Here, exploration means gathering more information about which workers are reliable, while exploitation is to select reliable workers based on the information at hand. To deal with this trade-off, the existing methods for worker selection adopt different strategies such as interval estimation. However, this is not enough in recent heterogeneous crowdsourcing where a worker may be reliable at only a subset of tasks with a certain type. For example, in named-entity recognition tasks for natural language processing, a worker may be good at recognizing the name of sports teams, but not be familiar with cosmetics brands. Therefore, it is more reasonable to model context-dependent reliability for workers in heterogeneous crowdsourcing. Here, context can be interpreted as the type for required skill of a certain task in heterogeneous crowdsourcing. In this thesis, we propose the bandit-based task assignment (BBTA) method, which is a contextual bandit formulation for worker selection in heterogeneous crowdsourcing. We also theoretically investigate BBTA and provide sublinear regret bounds, indicating that the performance of our worker selection strategy performs as well as the best one when the budget goes to infinity. Finally, we conduct extensive experiments on both benchmark data and real data, and the results demonstrate that BBTA can outperform other worker selection methods in heterogeneous crowdsourcing.

As the other important aspect of task-worker assignment, task selection focuses on selecting an appropriate task at each round. Although worker selection has been successfully addressed by the original BBTA, task selection has not been thoroughly investigated yet. A good strategy of task selection can also help us efficiently use the budget. In the original BBTA, we simply use a common uncertainty criterion for task selection. To investigate this problem more, we further develop several strategies for task selection and embed them into the original BBTA. The idea of these strategies is borrowed from query strategies in active learning. One of the proposed strategies called margin sampling is equivalent to the original one used in BBTA. We experimentally evaluate the proposed strategies for BBTA, and demonstrate that the performance of BBTA can be further improved by adopting appropriate task selection strategies such as the least confidence strategy.

Finally, we conclude that the proposed strategies of worker selection and task selection for crowdsourcing are successful and worth further investigating in the future.

備考 ： 論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).