Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 108C (2017) 285–294

Procedia Computer Science

International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland

Learning Robust Low-Rank Approximation for Crowdsourcing on Riemannian Manifold

Qian Li[1], Zhichao Wang[2]*, Gang Li[3], Yanan Cao[1], Gang Xiong[1], and Li Guo[1]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2] Tsinghua University, Beijing, China
[3] Deaking University, Melbourne, Australia

Abstract

Recently, crowdsourcing has attracted substantial research interest due to its efficiency in collecting labels for machine learning and computer vision tasks. This paper proposes a Riemannian manifold optimization algorithm, ROLA (*Robust Low-rank Approximation*), to aggregate the labels from a novel perspective. Specifically, a novel low-rank approximation model is proposed to capture underlying correlation among annotators meanwhile identify annotator-specific noise. More significantly, ROLA defines the label noise in crowdsourcing as *annotator-specific* noise, which can be well regularized by $l_{2,1}$-norm. The proposed ROLA can improve the aggregation performance when compared with state-of-the-art crowdsourcing methods.

Peer-review under responsibility of the scientific committee of the International Conference on Computational Science

Keywords: Crowdsourcing, Low-Rank, Riemannian Optimization

1 Introduction

Crowdsourcing is an efficient and inexpensive way to collect labelled data in many application domains, ranging from computer vision to natural language processing [9]. Services such as *Amazon Mechanical Turk* and *CrowdFlower* provide platforms where the requesters can post tasks and collect labels from online annotators. An advantage of crowdsourcing is that a large number of labels can be obtained in a short time with relatively low cost. Traditionally researchers resort to the redundancy mechanism for quality assurance, which assigns each question to multiple annotators and then aggregates their labels. Thus, a fundamental challenge arises in crowdsourcing: *how to infer true labels from noisy but redundant labels contributed by a crowd of unreliable annotators.*

In the past several years, numerous approaches based on latent statistical models have emerged and taken both the expertise of each annotator and the difficulty of questions into consideration. For instance, *Beta* distribution is chosen as the prior for characterizing the annotator's expertise in [8]. Moreover, some extensions further argue that the expertise of

*Corresponding author, email: wzchary@gmail.com.

1877-0509
Peer-review under responsibility of the scientific committee of the International Conference on Computational Science
10.1016/j.procs.2017.05.179

annotators varies across different types of questions, and they adopt the confusion matrix to model their biases and skills [3, 11, 6]. Besides the annotators' expertise, the difficulties of questions are also modeled in recent research work [18, 1]. For example, Whitehill [18] assumes the Gaussian prior distribution for both the annotators' capabilities and the question difficulties with different levels.

Although these latent statistical methods have been broadly applicable in many scenarios, some common limitations exist: first, it is unreasonable to simply assume that every label is unreliable. Considering the fact that labeling tasks are typically micro and simple to answer [9], annotators who are strongly motivated by interests or payment, tend to provide the correct labels for all questions. Neglecting this *annotator-specific* property may result in an inefficient and ineffective aggregation. Second, the noise is usually caused by various factors, and especially the carelessness or malicious behaviors may bring in arbitrary noise magnitude which is not sufficient to be modeled by traditional probabilistic distribution [14, 6, 7]. Third, each annotator in existing work is treated independently with individual label-generating model, while ignoring the underlying correlations among the annotators. More importantly, most statistic models for crowdsourcing involve the non-convex log-likelihood function, few studies provide theoretical justification on their convergence. Even though satisfactory performance of statistic models can be observed in practice, the aggregation results may not be global optimum.

Recent advances in low-rank approximation bring new insights into the underlying structure of the observed labels, and provide the potential for further improving the aggregation performance. This paper proposes an efficient and robust low-rank approach to infer the true labels from the noisy labels provided by crowdsourcing annotators. The main contributions of this paper are in the following aspects:

- This paper proposes a robust low-rank model for aggregating the crowd labels from a novel perspective. By defining the label noise as *annotator-specific* and sparse regularized by $l_{2,1}$-norm, the noise can be easily identified simplifying the later aggregation process.

- The probabilistic interpretation for the low-rank crowdsourcing model is also provided. The probabilistic inference validates the low-rank assumptions and provides the rigorous theory for the low-rank crowdsosurcing model.

- A novel optimization algorithm ROLA is developed to efficiently solve the proposed low-rank model. ROLA fully explores the matrix manifold of the observed aggregations and designs the gradient-based optimization to obtain the low-rank matrix for refined labels.

2 Related Work

Majority voting (MV) is one of the most popular crowd sourcing methods owing to its simplicity. However, their difference in expertise or motivations is usually ignored [22, 12]. Probabilistic approaches have been proposed to build statistical latent models that simulates the label-generating process to infer the true labels [3, 1]. *Dawid & Skene model* (DS) [3] models the annotator's expertise by a probabilistic confusion matrix. DARE [1] first evaluates the annotator's expertise via questionnaires with correct answers. Alternative methods evaluate the reliability of annotators by the spammer score, which indicates how spammer the annotator is [12]. Probabilistic matrix factorization (PMF) [6] was proposed to deduce the latent feature vectors for annotators and questions respectively, and then to complete unobserved labels for all questions. Then, the consensus judgment can be made using majority voting. Prior work [21]