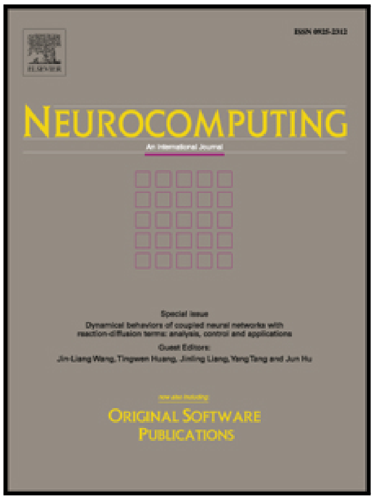
Communicated by Dr Bo Du

Accepted Manuscript



Multi-label Active Learning Based on Submodular Functions

Kuoliang Wu, Deng Cai, Xiaofei He

PII: S0925-2312(18)30707-0
DOI: 10.1016/j.neucom.2018.05.110
Reference: NEUCOM 19671

To appear in: *Neurocomputing*

Received date: 2 November 2017
Revised date: 5 April 2018
Accepted date: 29 May 2018

Please cite this article as: Kuoliang Wu, Deng Cai, Xiaofei He, Multi-label Active Learning Based on Submodular Functions, *Neurocomputing* (2018), doi: 10.1016/j.neucom.2018.05.110

Multi-label Active Learning Based on Submodular Functions

Kuoliang Wu, Deng Cai, Xiaofei He

State Key Lab of CAD&CG, Zhejiang University, No.388 Yu Hang Tang Road, Hangzhou 310058, China

Abstract

In the data collection task, it is more expensive to annotate the instance in multi-label learning problem, since each instance is associated with multiple labels. Therefore it is more important to adopt active learning method in multi-label learning to reduce the labeling cost. Recent researches indicate submodular function optimization works well on subset selection problem and provides theoretical performance guarantees while simultaneously retaining extremely fast optimization. In this paper, we propose a query strategy by constructing a submodular function for the selected instance-label pairs, which can measure and combine the informativeness and representativeness. Thus the active learning problem can be formulated as a submodular function maximization problem, which can be solved efficiently and effectively by a simple greedy lazy algorithm. Experimental results show that the proposed approach outperforms several state-of-the-art multi-label active learning methods.

Keywords: Multi-label Active Learning, Submodular Function Optimization

1. Introduction

In many supervised learning tasks, we need to collect plenty of training data and manually annotate. This procedure is very expensive due to the large amount of the data and the involvement of human experts. Active learning can automatically identify the most valuable data to query their labels, which can significantly reduce the amount of labeling effort needed to learn a given target function. There are two main query selection criteria for active learning. The first one is informativeness, which measures the uncertainty of the instance on the current model. The second criterion is representativeness. The instances selected by this criterion can well preserve the data distribution or its statistics. Active learning algorithms that deploy one of the two criteria are not sufficient to get the optimal result. As for informativeness, the queried data are not guaranteed to be i.i.d. sampled from the original data. And for representativeness, the methods based on this do not fully use the label information. Recent researches showed that methods combining these two criteria can result in better performance [12] [6] [22].

Traditional supervised learning problems assume that one instance is associated with only one single label. However, in many real world data analysis problem, each instance can be associated with more than one label [27]. For example, an image can be labeled as both *cat* and *dog* if these two animals appear simultaneously in the image; a piece of symphony can be labeled as *violin*, *classicalmusic* and *sentimental* if this is a sentimental classical music played by violin. Multi-label learning is the framework to handle such problem. In multi-label learning problem, the human annotator should decide whether one instance is relevant to every label when accessing the training data. Thus the labeling cost is much higher in multi-label learning than that of single label learning, which means the active learning query strategy is more necessary for multi-label learning. A lot of query strategies can be simply adapted from single-label active learning by transferring the multi-label task into a series of binary classification. For example, [15] proposed the method that selects the instance with the mean max loss and queries its labels. The method proposed in [25] aims to select instances leading to maximum loss reduction with the largest confidence. And the method named ADAPTIVE proposed in [14] takes both the uncertainty and label cardinality inconsistency into consideration. To handle the outlier labels for the measurement of uncertainty, a robust multi-label active learning algorithm [7] is proposed based on an MCC by merging uncertainty and representativeness. These multi-label active learning methods query the whole label vector for one instance in the unlabeled set. However, such approaches do not take into account the inherent relationships among the multiple labels. For example, if an image has the label *beach*, it is likely to have the label *sea*. So, if the classifier can catch and employ the label correlation to the multi-label learning, it's not necessary to query the whole labels of one instance.

Based on the fact mentioned above, several multi-label active learning methods are proposed to query the instance-label pairs in recent years. In each iteration of active learning procedure, these methods not only consider which instance to query, but also consider which label of the selected instance to query. Traditional multi-label active learning methods that query the whole label vector only make use of the information redundancy in sample dimension, and the methods that query the instance-label pairs can make use of information redundancy in both sample dimension and label dimension. Therefore they

Email addresses: wukuoliangzju@gmail.com (Kuoliang Wu), dengcai@gmail.com (Deng Cai), xiaofeihe@gmail.com (Xiaofei He)