Curriculum Learning of Multiple Tasks

Xuewen Yang

June 26 2018

Abstract

Sharing information between multiple tasks enables algorithms to achieve good generalization performance even from small amounts of training data. However, in a realistic scenario of multi-task learning not all tasks are equally related to each other, hence it could be advantageous to transfer information only between the most related tasks.

In this paper the author propose an approach that processes multiple tasks in a sequence with sharing between subsequent tasks instead of solving all tasks jointly. Subsequently,they address the question of curriculum learning of tasks, i.e. finding the best order of tasks to be learned.Their approach is based on a generalization bound criterion for choosing the task order that optimizes the average expected classification performance over all tasks.Their experimental results show that learning multiple related tasks sequentially can be more effective than learning them jointly, the order in which tasks are being solved affects the overall performance,and that their model is able to automatically discover a favourable order of tasks.

1. Introduction

Multi-task learning [1] studies the problem of solving several prediction tasks. While traditional machine learning algorithms can be applied to solve each task independently,they usually need significant amounts of labelled data to achieve generalization of reasonable quality. However,in many cases it is expensive and time consuming to annotate large amounts of data, especially in computer vision applications such as object categorization. An alternative approach is to share information between several related learning tasks and this has been shown experimentally to allow better generalization from fewer training points per task.

In this work the author focus on the parameter transfer approach to multi-task learning that rests on the idea that models corresponding to related tasks are similar to each other in terms of their parameter representations.They concentrate on the case of linear predictors and assume that similarity between the models is measured by the Euclidean distance between the corresponding weight vectors. In a multi-task setting this idea was introduced by Evgeniou and Pontil.There the authors propose an SVM-based algorithm that enforces the weight vectors corresponding to different tasks to lie close to some common prototype, and they show its effectiveness on several datasets. However, this algorithm treats all the tasks symmetrically, which might not be optimal in more realistic scenarios. There might be some outlier tasks or groups of tasks such that there is no similarity between the tasks from different groups. Hence, more flexible models are needed that are able to exploit the structure underlying tasks relations and avoid negative consequences of transferring information between unrelated tasks.

2. Related Work

While this paper is based on the idea of transferring information through weight vectors, other approaches to multitask learning have been proposed as well. A popular idea in the machine learning literature is that parameters of related tasks can be represented as linear combinations of a small number of common latent basis vectors. Argyriou emphet al. proposed a method to learn such representations using sparsity regularization. This method was later extended to allow partial overlap between groups of tasks. It was also adapted to the lifelong setting in [7], where Ruvolo and Eaton proposed a way to sequentially update the model as new tasks arrive, and discussed in [5], where a generalization bound for lifelong learning was first presented.

In [6], the model was extended to the case when the learner is allowed to choose which task to solve next and several heuristics were proposed for making this choice. Experimentally subspace-based methods have shown good performance in situations where many tasks are available and the underlying feature representations are low-dimensional.When the feature dimensionality gets larger, however, their computational cost grows quickly, and this makes them not applicable for the type of computer vision problems we are interested An exception is [3], where Jayaraman *et al.* apply subspace-based method to jointly learn multiple attribute predictors. However, even there, dimensionality reduction procedure was required.

Methods based on the sharing of weight vector have also been generalized since their original introduction, in particular to relax the assumption that all tasks have to be related.In [2], Evgeniou *et al.* achieved this by introducing a graph regularization. Alternatively, Chen *et al.* proposed to penalize deviations in weight vectors for highly correlated tasks. However, these methods require prior knowledge about the amount of similarities between tasks. In contrast,the algorithm we present in this work does not assume all tasks to be related, yet does not need a priori information regarding their similarities, either.

3. Experiments

In this section the author verify two main claims: 1) learning multiple tasks in a sequential manner can be more effective than learning them jointly; 2)finding automatically a favourable order in terms of average classification accuracy.The author use two publicly available datasets: Animals with Attributes (AwA) and Shoes augmented with attributes [4]. In the first experiment,The author study the case when each task has a certain level of difficulty for learning the object class, which is defined by human annotation in a range from easiest to hardest.They show the advantage of a curriculum learning model over learning multiple tasks jointly and learning each task independently.They also study the automatically determined orders in more details, comparing them with the orders when learning goes from easiest to hardest tasks in the spirit of human learning. In the second experiment,the author study the scenario of learning visual attributes that characterize shoes across different shoe models.In this setting, some tasks are clearly related such as high heel and shiny, and some tasks are not, such as high heel and sporty. Therefore,they also apply the variant of our algorithm that allows multiple subsequences, showing that it better captures the task structure and is therefore the favourable learning strategy.

3.1. Learning the order of easy and hard tasks

The author focus on eight classes from the AwA dataset: chimpanzee,giant panda, leopard, persian cat, hippopotamus,raccoon, rat, seal, for which human annotation is available,whether an object is easy or hard to recognize in an image. For each class the annotation specifies ranking scores of its images from easiest to hardest. To create easy-hard tasks,the author split the data in each class into five equal parts with respect to their easy-hard ranking and use these parts to create five tasks per class. Each part has on average 120 samples except the class rat.

As we can see from Figure 1, the proposed SeqMT method outperforms MT and IndSVM algorithms in all 8 cases. This shows that knowledge transfer between the tasks is clearly advantageous in this scenario, and it supports our claim that learning tasks sequentially is more effective than learning them jointly if not all tasks are equally related.

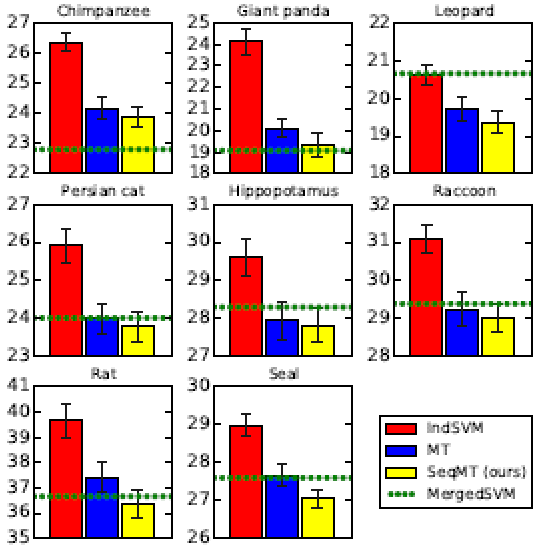

Figure 1. Learning the order of easy and hard tasks on AwA dataset: comparison of the proposed SeqMT method with the multi-task (MT) and the single-task (IndSVM) baselines. The height of the bar corresponds to the average error rate performance over 5 tasks across 20 repeats (the lower the better). As a reference, we also provide the MergedSVM baseline, trained on data that is merged from all tasks. For a complete table with all the results, please refer to the technical report [28].

References

[1] R. Caruana. Multitask learning. *Machine Learning*, 1997. 1

[2] T. Evgeniou, C. A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 2005. 2

[3] D. Jayaraman, F. Sha, and K. Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *CVPR*, 2014. 1

[4] A. Kovashka. Whittlesearch: Image search with relative attribute feedback. *IJCV*, 2015. 2

[5] A. Pentina and C. H. Lampert. A pac-bayesian bound for lifelong learning. In *ICML*, 2014. 1

[6] P. Ruvolo and E. Eaton. Active task selection for lifelong machine learning. In *AAAI*, 2013. 1

[7] P. Ruvolo and E. Eaton. ELLA: an efficient lifelong learning algorithm. In *ICML*, 2013. 1