Active Multitask Learning with Committees

Anonymous Authors[1]

Abstract

The cost of annotating training data has traditionally been a bottleneck for supervised learning approaches. The problem is further exacerbated when supervised learning is applied to a number of correlated tasks simultaneously since the amount of labels required scales with the number of tasks. To mitigate this concern, we propose an active multitask learning algorithm that achieves knowledge transfer between tasks. The approach forms a so-called *committee* for each task that jointly makes decisions and directly shares data across similar tasks. Our approach reduces the number of queries needed during training while maintaining high accuracy on test data. Empirical results on benchmark datasets show significant improvements on both accuracy and number of query requests.

1. Introduction

A triumph of machine learning is the ability to predict with high accuracy. However, for the dominant paradigm, which is supervised learning, the main bottleneck is the need to annotate data, namely, to obtain labeled training examples. The problem becomes more pronounced in applications and systems which require a high level of personalization, such as music recommenders, spam filters, etc. Several thousand labeled emails are usually sufficient for training a good spam filter for a particular user. However, in real world email systems, the number of registered users is potentially in the millions, and it might not be feasible to learn a highly personalized spam filter for each of them by getting several thousand labeled data points for each user.

One method to relieve the need of the prohibitively large amount of labeled data is to leverage the relationship between the tasks, especially by transferring relevant knowledge from information-rich tasks to information-poor ones, which is called multitask learning in the literature. We consider multitask learning in an online setting where the learner sees the data sequentially, which is more practical in real world applications. In this setting, the learner receives an example at each time round, along with its task identifier and then predicts its true label. Afterwards, the learner queries the true label and updates the model(s) accordingly.

The online multitask setting has received increasing attention in the machine learning community in recent years (Cavallanti et al., 2010; Abernethy et al., 2007; Dekel et al., 2007; Lugosi et al., 2009; Saha et al., 2011; Murugesan et al., 2016; Dekel et al., 2006). However, they make the assumption that the true label is readily available to be queried, which is impractical in many applications. Also, querying blindly can be inefficient when annotation is costly.

Active learning further reduces the work of the annotator by selectively requesting true labels from the oracles. Most approaches in active learning for sequential and stream-based problems adopt a measure of uncertainty / confidence of the learner in the current example (Cesa-Bianchi et al., 2006; Dekel et al., 2012; Orabona & Cesa-Bianchi, 2011; Cavallanti et al., 2009; Agarwal, 2013).

The recent work by Murugesan & Carbonell (2017) combines active learning with online multitask learning using *peers* or *related tasks*. When the classifier of the current task is not confident, it first queries its similar tasks before requesting a true label from the oracle, incurring a lower cost. Their learner gives priority to the current task by always checking its confidence first. In the case when the current task is confident, the opinions of its peers are ignored.

This paper proposes an active multitask learning framework which is more *humble*, in a sense that both the current task and its peers' predictions are considered simultaneously using a weighted sum. We have a *committee* which makes joint decisions for each task. In addition, after the true label of a training sample is obtained, this sample is shared directly to similar tasks, which makes training more efficient.

2. Problem Formulation

The problem formulation and setup are similar to (Murugesan et al., 2016; Murugesan & Carbonell, 2017). Suppose we are given K tasks and the k-th task is associated with

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

N_k training samples. We consider each task to be a linear binary classification problem, but the extensions to multi-class or non-linear cases are straightforward. We use the good-old perceptron-based update rule in which the model for a given task is only updated when the prediction for that training example is in error. The data for task k is $\{x_k^{(i)}, y_k^{(i)}\}_{i=1}^{N_k}$, where $x_k^{(i)} \in \mathbb{R}^D$ is the i-th instance from the k-th task, $y_k^{(i)} \in \{-1, +1\}$ is the corresponding label and D is the dimension of features. When the notation is clear from the context, we drop task index k and simply write $((x^{(i)},\, k),\, y^{(i)})$. We consider the online setting where the training example $((x^{(t)},\, k),\, y^{(t)})$ comes at round t.

Denote $\{w_k^{(t)}\}_{k\in[K]}$ the set of weights learned for the K binary classifiers at round t. Also denote $\boldsymbol{w} \in \mathbb{R}^{K\times D}$ the weight matrix whose k-th row is w_k. The label $\hat{y}^{(t)}$ is predicted based on the sign of the output value from the model. Then the hinge loss of task k on the sample $((x^{(t)},\, k),\, y^{(t)})$ at round t is given by $\ell_{kk}^{(t)} = \left(1 - y^{(t)}\langle x^{(t)}, w_k^{(t)}\rangle\right)_+$. In addition, we also consider the losses of its peer tasks m ($m \neq k$) as $\ell_{km}^{(t)} = \left(1 - y^{(t)}\langle x^{(t)}, w_m^{(t)}\rangle\right)_+$. $\ell_{km}^{(t)}$ indicates the loss incurred by using task m's knowledge / classifier to predict the label of task k's training sample. $\ell_{km}^{(t)}$ plays an important role in learning the similarities among tasks and hence the committee weights. Intuitively, two tasks should be more similar if one task's training samples can be correctly predicted using the other task's classifier.

The goal of this paper is to achieve a high accuracy on the test data, and at the same time to issue as small a number of queries to the oracle as possible during training, by efficiently sharing and transferring knowledge among similar tasks.

3. Active Multitask Learning w/ Committees

In this section we introduce our algorithm *Active Multitask Learning with Committees* (AMLC) as shown in Algorithm 1. This algorithm provides an efficient way for online multitask learning. Each task uses not only its own knowledge but also knowledge from other tasks, and shares training examples across similar tasks when necessary. The two main components of Algorithm 1 are described in Section 3.1 and 3.2. In Section 3.3, we compare AMLC with the state-of-the-art online multitask learning algorithm.

3.1. Learning with Joint Decisions

We maintain and update a *relationship matrix* $\boldsymbol{\tau} \in \mathbb{R}^{K\times K}$ through the learning process. The k-th row of $\boldsymbol{\tau}$, denoted τ_k, is the *committee weight vector* for task k, also referred to as *committee* for brevity. Element τ_{ij} of the relationship matrix indicates the closeness or similarity between task i and task j, and also the importance of task j in task i's committee in predicting. Given a sample $((x^{(t)},\, k),\, y^{(t)})$ at round t, the confidence of task k is jointly decided by its committee; namely, a weighted sum of confidences of all tasks, $p = \sum_{m\in[K]} p_{km}^{(t)}\tau_{km}^{(t-1)}$, where $p_{km}^{(t)} = \langle x^{(t)}, w_m^{(t-1)}\rangle$ for $m \in [K]$. Each confidence is just the common confidence measure for perceptron, using distance from the decision boundary (Cesa-Bianchi et al., 2006). The prediction is done by taking the sign of the confidence value. The learner then makes use of this confidence value by sampling a Bernoulli random variable, $P^{(t)}$, to decide whether to query the true label of this sample. The larger p is, the more likely for $P^{(t)}$ to be 0, signifying greater confidence. The hyperparameter b controls the level of confidence that the current task has to have to not request the true label.

Algorithm 1 Active Multitask Learning with Committees

function AMLC (b, C, T)
  Initialize $w_m^{(0)} = \mathbf{0}_D, \forall m \in [K], \boldsymbol{\tau}^{(0)} = \frac{1}{K}\mathbf{1}_{K\times K}$
  for $t = 1, 2, .., T$ **do**
    Receive $(x^{(t)}, k)$
    Compute $p_{km}^{(t)} = \langle x^{(t)}, w_m^{(t-1)}\rangle$ for $m \in [K]$
    $p = \sum_{m\in[K]} p_{km}^{(t)}\tau_{km}^{(t-1)}$
    Predict $\hat{y}^{(t)} = \text{sign}(p)$
    Draw $P^{(t)} \sim \text{Bernoulli}\left(\frac{b}{b+|p|}\right)$.
    if $P^{(t)} = 1$ **then**
      Query true label $y^{(t)}$ and set $M^{(t)} = 1$ if $y^{(t)} \neq \hat{y}^{(t)}$
      Update $w_k^{(t)} = w_k^{(t-1)} + P^{(t)}M^{(t)}y^{(t)}x^{(t)}$
      Update $\boldsymbol{\tau}$:

$$\tau_{km}^{(t)} = \frac{\tau_{km}^{(t-1)} e^{-C\cdot\frac{l_{km}^{(t)}}{\lambda}}}{\sum_{m'\in[K]} \tau_{km'}^{(t-1)} e^{-C\cdot\frac{l_{km'}^{(t)}}{\lambda}}}, m \in [K]$$

      for $\forall m \in [K]$, and $m \neq k$ **do**
        Set $S_m^{(t)} = 1$ if sign $\left(p_{km}^{(t)}\right) \neq \hat{y}^{(t)}$ and $\tau_{km}^{(t)} \geq \tau_{kk}^{(t)}$
        Update $w_m^{(t)} = w_m^{(t-1)} + S_m^{(t)}y^{(t)}x^{(t)}$
      end for
    end if
  end for
  return $\boldsymbol{\tau}^{(t)}\boldsymbol{w}^{(t)}$
end function

The learner only queries the true label when the current task's committee turns out to be unconfident. Another binary variable $M^{(t)}$ is set to be 1 if task k makes a mistake. Subsequently, its weight vector is updated following the conventional perceptron scheme. The learner then updates the relationship matrix following a similar policy as in (Murugesan et al., 2016; Murugesan & Carbonell, 2017). For tasks that incur no loss on this example, their weights in the committee are not changed. On the other hand, for tasks that incur non-zero loss, their weights are decreased by a factor $\exp(-C \cdot l_{km}^{(t)}/\lambda)$. The hyperparameter C decides

Models	Landmine Detection		Spam Detection		Sentiment	
	Accuracy	*#Queries*	*Accuracy*	*#Queries*	*Accuracy*	*#Queries*
Random	0.8914 ±0.0126	2323.5 ±11.5	0.7940 ±0.0131	751.8 ±14.3	0.6068 ±0.0065	1092.0 ±16.4
Independent	0.9070 ±0.0079	2770.3 ±25.8	0.8232 ±0.0159	1188.6 ±6.6	0.6404 ±0.0050	1987.5 ±7.1
PEER	0.9362 ±0.0025	1206.0 ±23.2	0.8334 ±0.0134	1085.7 ±13.9	0.6425 ±0.0067	1979.7 ±10.3
PEER+Share	0.9231 ±0.0112	1885.8 ±71.3	**0.8766 ±0.0135**	935.3 ±18.3	**0.7645 ±0.0077**	1754.5 ±9.2
AMLC	**0.9367 ±0.0020**	**189.1 ±12.0**	**0.8706 ±0.0102**	**321.3 ±15.6**	0.7358 ±0.0093	**633.9 ±11.4**

Table 1. Accuracy on test set and total number of queries during training over 10 random shuffles of the training examples. The 95% confidence level is provided after the average accuracy. The best performance is highlighted in bold. On Spam Detection, both PEER+Share and AMLC are highlighted because AMLC has a lower mean but also smaller variance.

how much decrease happens on the weight given non-zero loss, and $\lambda = \sum_{m=1}^{K} l_{km}^{(t)}$. These new weights are then normalized to sum to 1.

3.2. Data Sharing

To further encourage data sharing and information transfer between similar tasks, after the true label is obtained, the learner also shares the data with similar tasks of task k, so that peer tasks can learn from this sample as well. Similar tasks are identified by having a larger weight than the current task in the committee. We set $S_m^{(t)} = 1$ to indicate task m is a similar task to k and thus the data is shared with it.

3.3. Comparison with PEER

The most related work to ours is *active learning from peers* (PEER) from (Murugesan & Carbonell, 2017). In this section we discuss the main difference between our method and theirs with some intuition.

Firstly, we do not treat the task itself and its peer tasks separately. Instead, the final confidence of the current task is jointly decided using the confidences of all tasks, weighted by the committee weight vector. It is humble in a sense that it always considers its peer tasks' advice when making a decision. There are two main advantages of our approach. 1) For PEER, no updates happen and no knowledge is transferred when the current task itself is confident. This can result in difficulties for the learner to recover from being blindly confident. Blind confidence happens when the classifier makes mistakes on training examples but with high confidence, especially in early stage of training when data are not enough. 2) Our method updates the committee weight vector while keeping $\sum_{m\in[K]} \tau_{km} = 1$ instead of $\sum_{m\in[K],m\neq k} \tau_{km} = 1$. It then becomes possible that the current task itself has an equal or lower influence than other tasks on the final prediction. This is more desirable because identical tasks should have equal weights, and information-poor tasks should rely more on their information-rich peers when making predictions.

Secondly, our algorithm enables the sharing of training data across similar tasks directly, after acquiring the true label of this data. Querying can be costly, and the best way to make use of the expensive label information is to share it. Assuming that all tasks are identical, the most productive algorithm would merge all data to learn a single classifier. PEER is not able to achieve this because each task is still trained independently, since tasks only have access to their own data. Though PEER indirectly accesses others' data through querying peer tasks, this sharing mechanism can be insufficient when tasks are highly similar. In the case that all tasks are identical, our algorithm converges to a relationship matrix with identical elements and eventually all tasks are trained on every example that has been queried.

4. Experiments

In this section, we evaluate our proposed algorithm on three benchmark datasets for multitask learning, and compare our performance with many baseline models. We set $b = 1$ for all the experiments and tune the value of C from 20 values using 10-fold cross validation. Unless otherwise specified, all other model parameters are chosen via 10-fold cross validation.

4.1. Benchmark Datasets

Landmine Detection[1] consists of 19 tasks collected from different landmine fields. Each task is a binary classification problem: landmines (+) or clutter (-), and each example consists of 9 features. *Spam Detection*[2] consists of labeled training data: spam (+) or non-spam (-) from the inboxes of 15 users, and each user is considered as a single task. *Sentiment Analysis*[3] (Blitzer et al.) consists of product reviews from Amazon containing reviews from 22 domains. We consider each domain as a binary classification task:

[1] http://www.ee.duke.edu/~lcarin/LandmineData.zip

[2] http://ecmlpkdd2006.org/challenge.html

[3] http://www.cs.jhu.edu/~mdredze/datasets/sentiment/

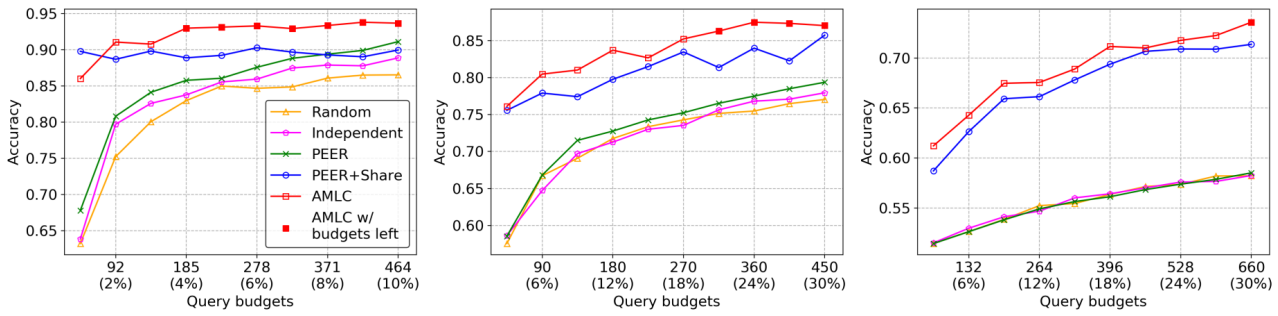

Figure 1. Accuracy on test set w.r.t. query budget. Left, middle and right are Landmine Detection, Spam Detection and Sentiment respectively. The filled markers indicate that there are still queries left in the budget.

positive review (+) and negative review (-). Details about our training and test sets are shown in Appendix A.

4.2. Results

We compare the performance of 5 different models. *Random* does not use any measure of confidence. Namely, $P^{(t)}$ is always a Bernoulli random variable with success probability 0.5. The probability of querying or not querying true label are equal. *Independent* uses the confidence which is purely computed form the weight vector of the current task. Obviously both Random and Independent have no knowledge transfer among tasks. *PEER* is the algorithm from (Murugesan & Carbonell, 2017). *AMLC* (Active Multitask Learning with Committees) is our proposed method as shown in Algorithm 1. In addition, we also show the performance of *PEER+Share*, in which we simply add to PEER the data sharing mechanism as illustrated in section 3.2.

Table 1 shows the accuracy on test set and the total number of queries (label requests) to oracles during training of five models. Each value is the average of 10 random shuffles of the training set. The 95% confidence level is also shown. Notice that our re-implementation of PEER achieves similar performance on the Landmine and Spam datasets but seems to perform worse on Sentiment. The reason is that we are using a different representation of the training examples. We use the default bag-of-words representation coming with the dataset and there are approximately $2900K$ features.

The highlighted values illustrate the best performance across all models. On Spam Detection, AMLC is also highlighted because it is more confident about its accuracy even though the actual value is slightly lower than PEER+Share. It can be seen that our proposed methods (PEER+Share and AMLC) significantly outperform the the others. PEER has better performance compared to Random and Independent but still behaves worse than PEER+Share and AMLC. It can be shown that simply adding data sharing can improve both accuracy and number of queries used during training. The only exception is on Landmine Detection, where PEER+Share requests more queries than PEER. Though simply adding data sharing results in improvement, after learning with joint decisions in AMLC, we observe further drastic decrease on the number of queries, while maintaining high accuracy.

Another goal of active multitask learning is to efficiently make use of the labels. In order to evaluate this, we give each model a fixed number of query budget and the training process is ended after the budget is exhausted. We show three plots (one for each dataset) in Figure 1. Based on the difficulty of learning from each dataset, we choose different budgets to evaluate (up to $10\%, 30\%$ and 30% of the total training examples for Landmine, Spam and Sentiment respectively). We can see that given a limited number of query budgets, AMLC outperforms all models on all three datasets, as a result of it encouraging more knowledge transfer among tasks. It is worth noting that the Landmine dataset is quite unbalanced (high proportion of negative labels), and PEER+Share and AMLC can achieve high accuracy with extremely limited number of queries. However, the classifier learned by PEER+Share is unconfident and thus it keeps requesting true labels in the following training process.

5. Conclusion

We propose a new active multitask learning algorithm that encourages more knowledge transfer among tasks compared to the state-of-the-art models, by using joint decision / prediction and directly sharing training examples with true labels among similar tasks. Our proposed methods achieve both higher accuracy and lower number of queries on three benchmark datasets for multitask learning problems. Future work includes theoretical analysis of the mistake bound and comparison with those of the baseline models. Another interesting direction is to handle unbalanced task data; in other words, one task has much more / less training data than the others.

References

Abernethy, J., Bartlett, P., and Rakhlin, A. Multitask learning with expert advice. In *International Conference on Computational Learning Theory*, pp. 484–498. Springer, 2007.

Agarwal, A. Selective sampling algorithms for cost-sensitive multiclass prediction. In *International Conference on Machine Learning*, pp. 1220–1228, 2013.

Blitzer, J., Dredze, M., and Pereira, F. Domain adaptation for sentiment classification. In *45th Annv. Meeting of the Assoc. Computational Linguistics (ACL07)*.

Cavallanti, G., Cesa-Bianchi, N., and Gentile, C. Linear classification and selective sampling under low noise conditions. In *Advances in Neural Information Processing Systems*, pp. 249–256, 2009.

Cavallanti, G., Cesa-Bianchi, N., and Gentile, C. Linear algorithms for online multitask classification. *Journal of Machine Learning Research*, 11(Oct):2901–2934, 2010.

Cesa-Bianchi, N., Gentile, C., and Zaniboni, L. Worst-case analysis of selective sampling for linear classification. *Journal of Machine Learning Research*, 7(Jul):1205–1230, 2006.

Dekel, O., Long, P. M., and Singer, Y. Online multitask learning. In *International Conference on Computational Learning Theory*, pp. 453–467. Springer, 2006.

Dekel, O., Long, P. M., and Singer, Y. Online learning of multiple tasks with a shared loss. *Journal of Machine Learning Research*, 8(Oct):2233–2264, 2007.

Dekel, O., Gentile, C., and Sridharan, K. Selective sampling and active learning from single and multiple teachers. *Journal of Machine Learning Research*, 13(Sep):2655–2697, 2012.

Lugosi, G., Papaspiliopoulos, O., and Stoltz, G. Online multi-task learning with hard constraints. *arXiv preprint arXiv:0902.3526*, 2009.

Murugesan, K. and Carbonell, J. Active learning from peers. In *Advances in Neural Information Processing Systems*, pp. 7011–7020, 2017.

Murugesan, K., Liu, H., Carbonell, J., and Yang, Y. Adaptive smoothed online multi-task learning. In *Advances in Neural Information Processing Systems*, pp. 4296–4304, 2016.

Orabona, F. and Cesa-Bianchi, N. Better algorithms for selective sampling. In *International conference on machine learning*, pp. 433–440. Omnipress, 2011.

Saha, A., Rai, P., DaumÃ, H., Venkatasubramanian, S., et al. Online learning of multiple tasks and their relationships. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 643–651, 2011.

Appendix

We describe in details about the three datasets and the train / test split used for the experiments. The description is adapted from (Murugesan et al., 2016; Murugesan & Carbonell, 2017).

A. Dataset Details

Landmine Detection consists of 19 tasks collected from different landmine fields. Each task is a binary classification problem: landmines (+) or clutter (-) and each example consists of 9 features extracted from radar images with four moment-based features, three correlation-based features, one energy ratio feature and a spatial variance feature. Landmine data is collected from two different terrains: tasks 1-10 are from highly foliated regions and tasks 11-19 are from desert regions; therefore tasks naturally form two clusters. Any hypothesis learned from a task should be able to utilize the information available from other tasks belonging to the same cluster.

Spam Detection is obtained from ECML PAKDD 2006 Discovery challenge for the spam detection task. We used the task B challenge dataset which consists of labeled training data from the inboxes of 15 users. We consider each user as a single task and the goal is to build a personalized spam filter for each user. Each task is a binary classification problem: spam (+) or non-spam (-) and each example consists of approximately $150K$ features representing term frequency of the word occurrences. Since some spam is universal to all users (e.g. financial scams), some messages might be useful to certain affinity groups, but spam to most others. Such adaptive behavior of users interests and dis-interests can be modeled efficiently by utilizing the data from other users to learn per-user model parameters.

Sentiment Analysis contains product reviews from many domains on Amazon. We consider each domain as a binary classification task. Reviews with rating > 3 are labeled positive (+), those with rating < 3 are labeled negative (-), and reviews with rating $= 3$ are discarded as the sentiments are ambiguous and hard to predict. We choose 22 domains which have enough data for both positive and negative labels. We use the default preprocessed bag-of-words representation that comes with the dataset and each example consists of approximately $2900K$ features.

We choose 3040 examples (160 training examples per task) for landmine, 1500 emails for spam (100 emails per user inbox) and 2200 reviews for sentiment (100 reviews per domain) for our experiments. For landmine and spam, we use the rest of the examples for test set, and for sentiment, we select another 300 (use all remaining data if not enough) examples for test set. On average, each task in landmine, spam, sentiment has 509, 400 and 395 examples respectively. Note that we intentionally kept the size of the training data small to drive the need for learning from other tasks, which diminishes as the training sets per task become large.