Variational Bayes Learning from Relevant Tasks Only

Jaakko Peltonen
Department of Information and Computer Science
Helsinki University of Technology
P.O. Box 5400, FI-02015 TKK, Finland
jaakko.peltonen@tkk.fi

Yusuf Yaslan
Department of Computer Engineering
Istanbul Technical University
34469 Maslak Istanbul, Turkey
yyaslan@itu.edu.tr

Samuel Kaski
Department of Information and Computer Science
Helsinki University of Technology
P.O. Box 5400, FI-02015 TKK, Finland
samuel.kaski@tkk.fi

Abstract

We extend our recent work on *relevant subtask learning*, a new variant of multi-task learning where the goal is to learn a good classifier for a task-of-interest with too few training samples, by exploiting "supplementary data" from several other tasks. It is crucial to model the uncertainty about which of the supplementary data samples are relevant for the task-of-interest, that is, which samples are classified in the same way as in the task-of-interest. We have shown that the problem can be solved by careful *mixture modeling*: all tasks are modeled as mixtures of relevant and irrelevant samples, and the model for irrelevant samples is flexible enough so that the relevant model only needs to explain the relevant data. Previously we used simple maximum likelihood learning; now we extend the method to variational Bayes inference more suitable for high-dimensional data. We compare the method experimentally to a recent multi-task learning method and two naive methods.

1 Introduction

In classification tasks there is often too little training data to estimate sufficiently powerful models. This problem occurs in bioinformatics, image classification, finding relevant texts etc. Possible solutions could be to restrict the classifier complexity by prior knowledge, or to gather more data. However, prior knowledge may be insufficient or may not exist, measuring new data may be too expensive, and there may not exist more samples of *representative* data. Most classifiers assume that learning data are representative, that is, they come from the same distribution as test data.

Often, *partially representative* data is available: in bioinformatics there are databases full of data measured for different tasks or under different conditions; for texts there is the web, and so on. They can be seen as training data from a (partly) different distribution as the test data. Suppose we have several sets that may contain some portion of relevant data; *can we use these partially relevant data sets to build a better classifier for the test data?*

This learning problem is a special *multi-task learning* [1] problem (in multi-task learning, learning a classifier for one data set is called a *task*). Our setting is different from traditional multi-task learning, where models have mainly been symmetrical and transfer to new tasks is done by using the posterior from other tasks as a prior (e.g. [2]). By contrast, our problem is fundamentally asymmetric and more structured: test data fits one task, the "*task-of-interest*," and other tasks may contain *subtasks* relevant for the task-of-interest, but no other task needs to be wholly relevant.

Recently we provided a probabilistic solution to the problem based on a simple multitask mixture modeling approach called *relevant subtask learning* (RSL; [3]). Our models are better suited in the task-of-interest yet have the same order of complexity as earlier multi-task models. The most closely related model is the recent work in [4] which essentially learns a weighting (density ratio) for supplementary data samples; our approach is different in that we find relevant supplementary samples by modeling whether they can be classified with the same classifier, rather than by modeling density ratios; another difference is that our model learns simultaneously but separately from each task (supplementary data set). In this paper we extend our method to variational Bayes inference.

2 Relevant Subtask Learning

Consider a set of classification tasks indexed by $S = 1, ..., M$, with a small training data set available for each task. Each training data sample $(\mathbf{x}, c)$ has features $\mathbf{x}$ and class label c; for simplicity each task has here two classes. One task, with index U, is the *task-of-interest*: we want to perform well on this task, and future test data will come from the distribution of this task. The other tasks S are *supplementary* tasks. We will build a probabilistic class prediction model for the task-of-interest and the supplementary tasks. We introduce our solution by applying it to a simple parametric model; it can easily be generalized to more general parametric or semiparametric models.

Let $p(c|\mathbf{x}, U; \boldsymbol{\theta})$ be our model for the task-of-interest U, where $\boldsymbol{\theta}$ denotes all parameters. For each supplementary task S we assume that part of the samples come from the same distribution $p(c|\mathbf{x}, U; \boldsymbol{\theta})$ as the task-of-interest; these samples are relevant to us. The rest of the samples are not relevant to us; they come from a different distribution $p_{\text{nonrelevant}}(c|\mathbf{x}, S; \boldsymbol{\theta})$ specific to task S. Since each supplementary task S is a mix of relevant and nonrelevant data, we model it by a mixture model:

$$p(c|\mathbf{x}, S; \boldsymbol{\theta}) = (1 - \pi_S)p(c|\mathbf{x}, U; \boldsymbol{\theta}) + \pi_S p_{\text{nonrelevant}}(c|\mathbf{x}, S; \boldsymbol{\theta}) \quad (1)$$

where the mixing coefficients π_S can be different for each supplementary task S. In this paper we model the task-of-interest with logistic regression, and each of the nonrelevant distributions also with logistic regression, yielding $p(c|\mathbf{x}, U; \boldsymbol{\theta}) = (1 + \exp(-c\mathbf{w}_U^T\mathbf{x}))^{-1}$ and

$$p(c|\mathbf{x}, S; \boldsymbol{\theta}) = (1 - \pi_S)/(1 + \exp(-c\mathbf{w}_U^T\mathbf{x})) + \pi_S/(1 + \exp(-c\mathbf{w}_S^T\mathbf{x})) \quad (2)$$

where $\mathbf{w}_U$, the $\mathbf{w}_S$, and the π_S are parameters of our model.[1] Fitting this multitask mixture model to data extracts the useful parts from each supplementary task.

Previously we fitted the model by simple maximum likelihood learning [3], which is not well suited for high dimensional data; in this paper we introduce variational Bayes inference, where uncertainty of parameter values is taken into account. In brief, we use a mean field variational approximation for the posterior, where component distributions are in the exponential family (Gaussians for $\mathbf{w}_U$ and the $\mathbf{w}_S$, Beta for the π_S, Bernoulli for latent variables). Two noteworthy details are: first, because our model is a mixture model we introduce latent variables telling which mixture component generated each data point, and optimize their posterior as well; second, because the logistic functions in the likelihood are not in the exponential family, during optimization we use an exponential-family approximation for them which is updated at each step. The computational cost at each iteration is linear with respect to the number of samples. We call this algorithm vb-RSL.

3 Experiments

We compare variational Bayes based relevant subtask learning (vb-RSL) against three comparison methods, on a continuum of toy data domains and on a more realistic text classification task.

All comparison methods use variational Bayes inference. The main comparison method is a hierarchical model called *symmetric multitask learning* (SMTL; [5]); it assumes each task (data set) can be modeled by one logistic regression model, and that tasks arise from task clusters, where a single logistic regression model is used for all tasks inside each cluster. We used an implementation provided by the authors of [5]. We also use two naive methods: in *single-task learning* (vb-STL) supplementary data is ignored, and a logistic regression model is learned for the task-of-interest from its own

[1] As usual, we include the bias in the weights ($\mathbf{w}_U$ and the $\mathbf{w}_S$), yielding standard logistic regression when one element in the inputs $\mathbf{x}$ is constant.

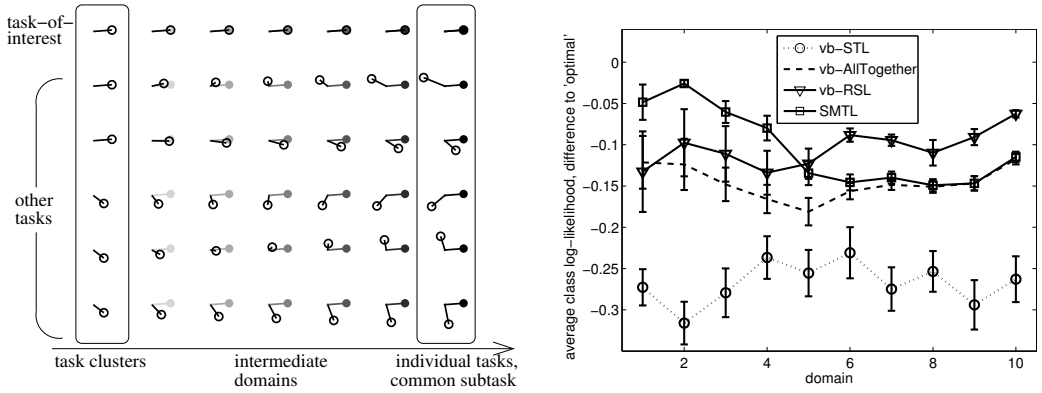

Figure 1: Performances on a continuum of toy multitask problem domains. **Left**: conceptual illustration; the lines in each glyph represent logistic regression classification directions used to generate one data set. In each domain (column) the topmost task is the task-of-interest. At left tasks are ideal for SMTL: they come from clusters, with a single logistic regression per cluster. At right tasks are ideal for vb-RSL; they all have an individual component and a shared component. Gray shades denote the mixing proportion of the shared component, which vanishes at left. **Bottom**: Results, averaged over 30 multitask problems for each domain. Numbers are performance differences (difference of average class log-likelihood) compared to approximately optimal results; lines show the average difference on each domain, and error bars show standard deviation of the mean. At left SMTL performs best, and at right vb-RSL performs best, matching the design of the continuum; naive methods perform poorly as expected.

data set only. In "*all together*" (vb-AllTogether) all supplementary data are assumed relevant for the task-of-interest, and a single logistic regression model is learned for the task-of-interest from all data pooled together. The naive methods are special cases of vb-RSL, where prior relevance probabilities of supplementary data samples are set to zero or one, respectively.

The first experiment is on a *continuum of multi-task classification problem domains*. The aim is to show that both vb-RSL and SMTL have domains where they work well; it is up to the analyst to decide which model is best for each application, based on experiments or domain knowledge.

Figure 1 (left) illustrates the continuum; at left, tasks (data sets) match the assumptions of SMTL: tasks come from clusters where one logistic regression model suffices to classify the data. At right, tasks match the assumptions of vb-RSL: all tasks are unique, but each contains some proportion of data relevant to the task-of-interest. We ran the methods on 30 multitask classification problems from each domain. We evaluate all methods by class log-likelihood, on test data from the task-of-interest. We also know an approximately optimal model for each toy problem; we show difference of results compared to the approximately optimal models. Figure 1 (right) shows the results: as expected, SMTL and vb-RSL work well for their own domains and the naive methods perform badly.

In the second experiment, the task is to predict *relevance of Reuters news articles* for a simulated "user-of-interest", given a set of labeled articles from that user and several sets of labeled articles from other users. The user-of-interest consider articles from a particular underlying news category to be relevant; the other users want such articles only part of the time, and at other times consider another category relevant. We generated 10 repetitions of this problem and ran all methods for each repetition, evaluating their performance on left-out news articles from the user-of-interest.

We found that performance of the methods depends on the domain parameters of the problem (dimensionality, number of tasks, amount of data, and proportion of relevant data). There seems to be a region of domain parameters where vb-RSL performs well but outside the region performance drops. For brevity we show only part of our results: Figure 2 shows performance as a function of the numbers of dimensions and data samples. The vb-RSL method gives the best results when there are many dimensions but few samples per data set (less than 100), which is a realistic scenario. SMTL is not much better than naive "all together" on this data. Results of vb-RSL improve when dimensionality grows. Interestingly, vb-RSL worsens when the number of samples is too high: this may

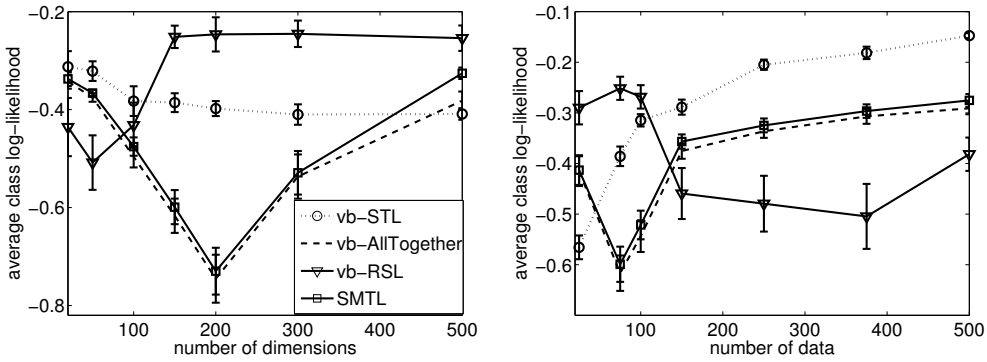

Figure 2: Comparison of vb-RSL to SMTL and two naive methods on Reuters data. The "default settings" of the multitask problem domain are 150 input dimensions, 10 tasks (data sets), 75 samples per data set, where the relevance proportion (proportion of samples that are relevant for the task-of-interest) is 0.5. In each subfigure we vary one of these parameters and keep the others fixed. The curves show test-set class prediction performance; the error bars show standard deviation of performance across 10 problems. **Left**: results as a function of the data dimensionality. **Right**: results as a function of the number of data samples per data set.

be because the number of latent variables in the RSL mixture model grows with data, and with too many variables the factorized variational posterior no longer approximates the real posterior well.

4 Conclusions

We introduced a variational Bayes method for the new problem of *relevant subtask learning* (RSL), where multiple supplementary tasks are used to learn one task-of-interest. Our method, vb-RSL, fits a multitask mixture model, where relevant data are modeled by a shared mixture component and nonrelevant data by components specific to each supplementary task.

Our approach is a viable alternative to the more traditional hierarchical modeling-based multitask learning. Comparing our new method experimentally to a recent multitask method (SMTL) and naive methods, we found that the our method and SMTL both have their own domains where they do well. It is up to the analyst to decide which method to use for each application. In further work we will improve the inference and investigate which applications are best suited for the RSL approach.

Acknowledgments

J. Peltonen and S. Kaski belong to Helsinki Institute for Information Technology HIIT and to the Adaptive Informatics Research Centre. The work was supported by the Academy of Finland, decision 207467. Y. Yaslan was supported by the Center for International Mobility CIMO. This work was also supported in part by the PASCAL2 Network of Excellence of the European Community. This publication only reflects the authors' views. We thank Y. Xue for providing SMTL code.

References

[1] R. Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

[2] R. Raina, A. Y. Ng, and D. Koller. Transfer learning by constructing informative priors. In *Inductive Transfer: 10 Years Later, NIPS 2005 workshop*. 2005.

[3] S. Kaski and J. Peltonen. Learning from relevant tasks only. In *Machine Learning: ECML 2007*, pages 608–615. Springer, 2007.

[4] S. Bickel, J. Bogojeska, T. Lengauer, and T. Scheffer. Multi-task learning for HIV therapy screening. In *proceedings of ICML 2008*, pages 56–63. Omnipress, 2008.

[5] Y. Xue, X. Liao, L. Carin, and B. Krishnapuram. Multi-task learning for classification with Dirichlet process priors. *Journal of Machine Learning Research*, 8:35–63, 2007.