Adaptive Information Filtering with Minimal Instruction MIT2000-08

Progress Report: July 1, 2002—December 31, 2002

Tommi S. Jaakkola and Tomaso Poggio

Project Overview

This project concerns with automated methods for finding a few pieces of relevant information (such as research articles) within a large dataset of predominantly incomplete and superficially similar information (such as technical report archives). While many such information filtering tasks vary considerably depending on the context, the primary challenges associated with automated techniques are often shared across different tasks. In this project, we develop the foundations for adaptive information retrieval tools with the ability to function accurately with minimal instruction of what is relevant, learn from related filtering problems, and make use of any optional feed-back automatically queried from the user

Progress Through December 2002

We have made progress on several fronts and outline here three areas in particular: 1) classification with unlabeled examples, 2) collaborative filtering, and 3) feature induction for text classification.

Classification with unlabeled examples. We have developed a new principle for combing unlabeled and labeled examples for the purpose of accurate classification of documents. The key motivation for this work is that the labels or relevance assessments that one is interested in learning to predict are often largely missing in the available data. Standard statistical methods are therefore not effective in this context as they have been developed primarily under the assumption that the key information is missing in only a small fraction of the available cases. Our principle is the first to articulate in probabilistic terms how the conditional distribution of labels given documents -- quantity needed for classification -- should in general relate to the marginal distribution of documents or the large number of unannoted documents. The principle is cast in terms of regularization theory and penalizes information about the labels that is introduced beyond the few available labeled examples. This ensures that the classification decisions for unlabeled documents are based firmly on the available information. Effectively, the regularization principle states how the conditional probabilities should be interpolated over the unlabeled points on the basis of the few labeled examples. More generally, the principle is formulated in terms of Tikhonov regularization theory.

(While information regularization was already mentioned in an earlier report, the idea has been extended considerably in the last six months.)

We are actively developing efficient algorithms for exploiting this principle in practical applications. We currently have two publications pertaining to early versions of the information regularization work:

> Ph.D. thesis by Martin Szummer (jointly advised by T. Jaakkola and T. Poggio):
> Final version of the paper presented at the NIPS conference:

Other material currently in preparation will be distributed as they are completed.

Collaborative filtering. Collaborative filtering provides a simple and efficient way of exploiting what is common to groups of decision makers. A typical collaborative filtering approach involves finding a low rank factorization of the data matrix containing user identities and their preference decisions. The low rank factorization that should capture shared features across users is, however, often estimated by assuming a simple squared loss between the data matrix and the predictions. This error measure is mathematically convenient but rarely reflects the structure of the problem. More appropriate error measures such as those based on probabilistic classification methods make the low rank factorization problem somewhat more challenging from a computational perspective. In particular, we can no longer obtain the solution in closed form and have to find the underlying low rank representation with iterative algorithms. We have developed several new algorithms for finding low rank matrix factorizations in the more general contexts relevant for collaborative filtering. The simplest such algorithms cast the iterative steps of the estimation process in terms of incomplete data estimation and formulate an EM algorithm for finding the low rank matrix factorizations. The mathematical foundation of these ideas can be found in the following technical report

N. Srebro and T. Jaakkola, "Generalized Low-Rank Approximations", AI Memo 2003-001.

The paper detailing an application of these ideas to collaborative filtering is under preparation and will be made available shortly.

Feature induction for text classification. An important problem in document classification is to find the set of text features that are useful for classification, both in general and in the context of a specific classification task. We cast the problem in terms of data compression and systematically search for features in the documents so as to best compress the labels (or summaries) that one is interested in predicting. The features come in two forms: those that are common to a number of classification tasks, and those that are useful only within a single task. Generic features that are useful across tasks are much less costly to introduce (they need to be encoded only once) but may not suffice for the specific task at hand.

A brief and preliminary description of some of these ideas can be found in:

J. Rennie and T. Jaakkola, “Feature induction for text classification”, AI Abstract 2003,

Research Plan for the Next Six Months

Our plan for the next six months consists primarily of the following tasks:

Developing and testing efficient algorithms for exploiting information regularization principle in large scale applied problems. Specifically, we hope to solve the regularization problem within a constrained class of conditional probabilities.

More extensive adaptation of generalized low rank estimation algorithms to collaborative filtering tasks.
Formulating and testing active learning methods in combination with the information regularization framework. This approach would not only make an appropriate use of the available unlabeled documents but would also exploit resulting classification decisions to iteratively find the most informative new documents to label.

We seek to extend the overall active learning framework to efficiently annotate documents in large databases with multiple topic labels. This would be done in collaboration with Dr. Naonori Ueda.