Stable Mixing of Complete and Incomplete Information

Adrian Corduneanu & Tommi Jaakkola

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: An increasing number of parameter estimation tasks involve the use of at least two information sources, one complete but limited, the other abundant but incomplete. Standard algorithms such as EM (or em) used in this context are unfortunately not stable in the sense that they can lead to a dramatic loss of accuracy with the inclusion of incomplete observations. Although stability can be achieved by downweighting the effect of incomplete data at the expense of smaller potential gains, standard algorithms do not offer any guidance for determining the optimal weighting.

Motivation: Many modern application areas such as text classification involve estimating generative probability models under limited labeled and abundant unlabeled data. Empirically [5] unlabeled data provides under model constraints rich information valuable for classification, thus its inclusion in training may lead to a significant increase in accuracy. However, unlabeled data alone cannot identify the assignment of labels to classes; trusting it too much may actually hurt performance. Indeed, experiments show that in a standard maximum likelihood setting, the inclusion of unlabeled data may dramatically improve as well as degrade the accuracy. It is imperative to find methods that remain stable while fully exploiting the potential of incomplete information.

Previous Work: Many algorithms for complete-data only or incomplete-data only estimation have been adapted to combine the two sources of information, most of them relying on heuristics without strong theoretical justification. EM [4] is the most popular, and it has been successfully applied in domains such as text classification [5]. Extensions of this idea include Co-training [2] which can be shown to work well at least under rather strong modeling assumptions, and kernel expansions [6] that change the problem by means of non-parametric density estimates. Castelli [3] gives a theoretical argument under unrealistic assumptions showing that incomplete information is exponentially less important then complete information. However, we are not aware of any previous analysis of the stability of such algorithms. Our discussion here is limited to approaches based on generative models.

Approach: We start from a criterion that explicates the mixing λ between complete and incomplete information, optimized by a standard iterative algorithm such as EM or Amari's em [1]. Instead of finding the typical fixed point for a given λ, we cast the problem in terms of differential equations that govern locally optimal solutions as a function of λ and we continuously evolve such solutions. We start from the unique complete-data solution at $\lambda = 0$, and increase the mixing towards incomplete-data only estimation specified by $\lambda = 1$. The intuition here is that models whose parameters can be continuously traced back to the complete-data solution are both well-grounded in the available complete information and make a good use of the incomplete data.

The advantage of our approach is that we can explicitly identify critical λ's after which the solution of the differential equation cannot be extended continuously. Evolving fixed points beyond such critical mixing would break the connection with complete information (the choice of the ensuing paths not determined by the complete information) leading to unpredictable classification performance. Our method finds the best mixing of the data sources (largest λ) that still leads to a stable solution. The resulting locally optimal density esitmate is likely to differ from that provided by, e.g., the EM algorithm for the same value of the mixing parameter λ.

To illustrate our method we applied it to the *20 newsgroups* document classification task. The generative model was naive Bayes with binary word occurrence features, and the criterion was weighted log-likelihood of both labeled and unlabeled data optimized by standard EM. You can see in Figure 1 that accuracy drops significantly at critical mixing, and that our algorithm, DIFFEM, is more stable than EM.

Impact: The proposed algorithm leads to stable estimation when standard algorithms would degrade performance by potentially over-emphasizing incomplete information. The idea of tracing paths of solutions via differen-

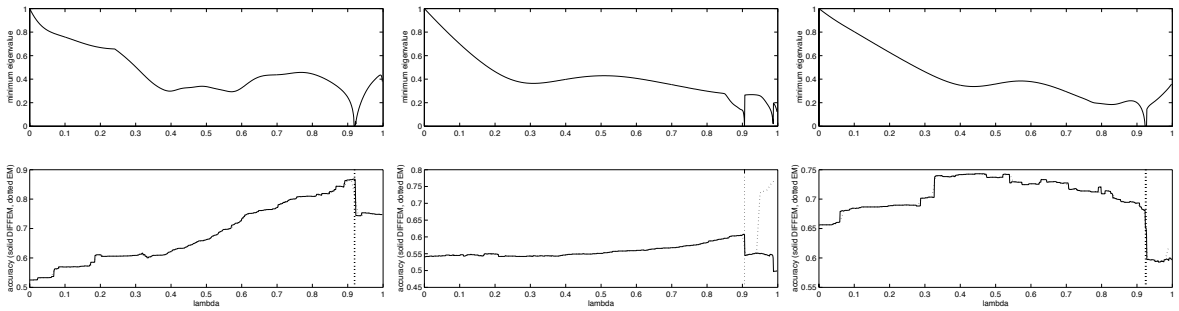

Figure 1: Three runs of DIFFEM versus EM on 50 labeled and 5000 unlabeled samples. For each run the upper graph is a criterion that signals a critical λ when 0, and the lower graph the classification accuracy of DIFFEM and EM as a function of λ. Most of the time the dotted-line EM coincides with DIFFEM.

tial equations is very general and has potential to improve other algorithms that undergo phase transitions as the source allocation, temperature, or other parameter is varied. This work also demonstrates the utility of geometric methods in solving fundamental estimation problems.

Future Work: Since the exact solution of the differential equation can be found in $O(n^3)$ time in the number of parameters, more efficient approximate methods or methods for better exploiting the structure of the problem need to be developed. Also, it is important to be able to evolve solutions beyond critical mixing by identifying all branches of fixed points that can follow.

Research Support: This work was supported in part by Nippon Telegraph and Telephone Corporation, by ARO MURI grant DAAD19-00-1-0466, and by NSF ITR grant #IIS-0085836.

References:

[1] S. Amari. Information geometry of the EM and em algorithms for neural networks. *Neural Networks*, 8(9):1379–1408, 1995.

[2] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, 1998.

[3] V. Castelli and T. M. Cover. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on Information Theory*, 42(6):2102–2117, November 1996.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–22, 1977.

[5] K. Nigam, A.K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39:103–134, 2000.

[6] M. Szummer and T. Jaakkola. Kernel expansions with unlabeled examples. In *Advances in Neural Information Processing Systems*, volume 13, pages 626–632, 2000.