

Semi-supervised or Semi-unsupervised?

Hal Daumé III
School of Computing
University of Utah
Salt Lake City, UT 84112
me@hal13.name

1 Some Definitions

We are interested in learning something using both labeled and unlabeled data, or else we wouldn't be at this workshop. The question I'd like to think about is: why do we want to do this? Is it:

1. because we think that adding a little labeled data to our pile of unlabeled data will help; *or*
2. because we think that adding a little unlabeled data to our pile of labeled data will help?

Typical approaches in NLP to the unlabeled+labeled problem fall into one of these two categories. In the first case, we basically have some unsupervised learning system that we know does fairly well on its own, and we're adding labeled data just to help it tweak things in a slightly better way. In the second case, we basically have some supervised learning system that we know does fairly well on its own, and we're adding unlabeled data to allow it to get a better sense of what real data "looks like." A good example of the first case that I know first hand is LEAF (Fraser and Marcu, 2007), since Alex Fraser was my office-mate for many years. A good example of the second case is graph-based methods for sentiment analysis (GBM) (Goldberg and Zhu, 2006), chosen since Andrew will be here!

A caricature of these results is as follows:

LEAF: Build a generative model explaining word alignments for machine translation that works in an unsupervised manner. Add in a little labeled data to tweak a dozen parameters.

GBM: Build a hyperplane-based sentiment classification algorithm and add in a bunch of unlabeled data to help find a better hyperplane.

Note that these work in entirely different realms. If we gave LEAF only unlabeled data, it would work okay; if we gave it only labeled data, it would work terribly. The reverse is true for GBM: with only unlabeled data, it would flounder; with only labeled data, it would do reasonably well.

Let's call¹:

1. little labeled = "semi-unsupervised"; *and*
2. little unlabeled = "semi-supervised"

(Although I've said "little", the quantity doesn't matter: what matters is whether we expect the systems to do well with only one part of the data.)

2 Why Does it Matter?

There is a strong correlation between semi-unsupervised and generative modeling; and a growing correlation between semi-supervised and discriminative modeling ("hybrid" approaches notwithstanding; more on these later). Generative models, and their even more shiny Bayesian counterparts, have grown to be the framework of choice for unsupervised learning. Since semi-unsupervised models are basically unsupervised models tweaked by some labeled data, it's then not surprising that this framework has a bit of a choke-hold on semi-unsupervised learning. On the other hand, since everyone agrees that discriminative modeling is better when you have labeled data, methods that try to use unlabeled data to improve discriminative approaches in the semi-supervised case are natural. This brings us to two alternatives, which I think are interesting to consider in the future, and which provide me ample opportunity for self-citation.

¹This is in analogy to semi-formal attire, which is almost formal attire but not quite.

2.1 “Hybrid” Approaches

A recent flurry of work (Bouchard and Triggs, 2004; Lasserre et al., 2006; Bouchard, 2007; Druck et al., 2007; Fujino et al., 2007; Agarwal and Daumé III, 2009) combines generative and discriminative models either by interpolation (not so interesting) or regularization (more interesting). A related piece of work that’s often left off, but is very interesting, is that of Suzuki and Isozaki (2008) that does something very similar to these hybrid approaches, but combining HMMs and CRFs, rather than the more boring naive Bayes and logistic regression. The interesting future direction I see here is using good generative models, rather than stupid things like naive Bayes. The Suzuki and Isozaki work is a great step in this direction. I hope we see a lot more.

2.2 Non-generative Unsupervised Learning

Shockingly, it’s possible to do unsupervised learning in a not-explicitly-generative fashion. Yes, I know, for years I too thought that it wasn’t. That’s why I wasted five years of my life learning about graphical models and Bayes stuff. But check out ICML this year: John, Rus and Tong have a paper that essentially replaces the generative mumbo-jumbo in dynamic models (think Kalman filters) with classifiers (Langford et al., 2009); I have a paper that uses the idea of self-prediction to do unsupervised learning for structured prediction using classifiers (Daumé III, 2009), called “Unsearn.” The thing that I think is cool about unsearn is that it shows that unsupervised learning doesn’t have to mean making feature independence assumptions and training things generatively. You could use SVMs or decision trees or whatever as your base learners and still do something reasonable. It also very naturally works in *either* a semi-supervised or semi-unsupervised fashion: there’s not really a hard line that forces it one way or the other. (Plus, semi-supervised learning in unsearn works *remarkably* well.)

3 The End

At the end of the day, machine learning is about getting knowledge into a system. If you don’t have many labels, you’d better have some strong priors. If you have lots of labels, you can forgo the priors. But let’s not: let’s do both and build the best systems anyone has ever seen.

References

- Arvind Agarwal and Hal Daumé III. 2009. Exponential family hybrid semi-supervised learning. In *International Joint Conference on Artificial Intelligence*, Pasadena, CA.
- Guillaume Bouchard and Bill Triggs. 2004. The trade-off between generative and discriminative classifiers. In *IASC International Symposium on Computational Statistics (COMPSTAT)*.
- Guillaume Bouchard. 2007. Bias-variance tradeoff in hybrid generative-discriminative models. In *ICMLA '07: Proceedings of the Sixth International Conference on Machine Learning and Applications*.
- Hal Daumé III. 2009. Unsupervised search-based structured prediction. In *International Conference on Machine Learning*, Montreal, Canada.
- Gregory Druck, Chris Pal, Andrew McCallum, and Xiaojin Zhu. 2007. Semi-supervised classification with hybrid generative/discriminative methods. In *Conference on Knowledge Discovery and Data Mining (KDD)*.
- Alexander Fraser and Daniel Marcu. 2007. Getting the structure right for word alignment: LEAF. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Akinori Fujino, Naonori Ueda, and Kazumi Saito. 2007. A hybrid generative/discriminative approach to text classification with additional information. *Inf. Process. Manage.*, 43(2).
- Andrew Goldberg and Xiaojin Zhu. 2006. Seeing stars when there aren’t many stars: Graph-based semi-supervised learning for sentiment categorization. In *HLT-NAACL Workshop on Textgraphs: Graph-based Algorithms for Natural Language Processing*.
- John Langford, Ruslan Salakhutdinov, and Tong Zhang. 2009. Learning nonlinear dynamic models. In *ICML*.
- Julia Lasserre, Christopher Bishop, and Thomas Minka. 2006. Principled hybrids of generative and discriminative models. In *Computer Vision and Pattern Recognition (CVPR)*.
- J. Suzuki and H. Isozaki. 2008. Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled data. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.