Semi-supervised learning and model search

Fabio Gagliardi Cozman FGCOZMAN@USP.BR
Escola Politécnica, University of São Paulo, Av. Prof. Mello Moraes, 2231 - Cidade Universitária 05508900, São Paulo, SP - Brazil

Ira Cohen IRACOHEN@IFP.UIUC.EDU
The Beckman Institute, 405 N. Mathews Ave., Urbana, IL 61801

Marcelo Cesar Cirelo MARCELO.CIRELO@POLI.USP.BR
Escola Politécnica, University of São Paulo

1. Introduction

The purpose of this debate paper is to pose the following question: Is it possible to develop general methods that handle semi-supervised learning without conducting some form of model search? We conjecture that the answer to this question, once a proper context is put in place, is *no*.

2. Semi-supervised learning

We assume that a database containing labeled and unlabeled data is given, and that a classifier is learned using the database. A classifier is a function from the *attributes* $\mathbf{X}$ in the database to a particular variable called the *class* variable C. A general strategy is to derive a classification rule from an estimate of the probability density $p(C, \mathbf{X})$ — as the exact form of this density is generally unknown, a *model* for the density is postulated. Another strategy is to construct the best classification rule that tries to minimize classification error without any attempt to construct a probabilistic model for C and $\mathbf{X}$. Classifiers that follow the first strategy often fall under the banner of *generative* methods (for example, the Naive Bayes classifier), while classifiers that follow the second strategy often fall under the banner of *diagnostic* methods (for example, classification trees and support vector machines).

Recent work on semi-supervised learning using generative models has met considerable success by using simple classifiers and employing the EM algorithm for maximum likelihood estimation [9]. There has also been notable theoretical analysis indicating that unlabeled data can be very useful [3, 10]. However, these optimistic results cannot be easily reproduced across a broad spectrum of learning problems; simulation studies and experiments with real databases have revealed that unlabeled data can actually be detrimental to classification performance [4, 5]. Even reports of successful application of semi-supervised learning contain situations where unlabeled data leads to performance degradation [1, 2, 9, 11]. In short, there are many situations where the addition of more unlabeled samples to a learning procedure consistently leads to classifiers with larger classification error.

There have been conjectures in the literature concerning the reasons for such failure of semi-supervised learning. Typically these explanations postulate serious modeling problems, such as the existence of classes that are never measured in the labeled data, or differences between the distribution for labeled and for unlabeled data [9, 11]. Two issues raised at the call-for-papers for the workshop are illustrative: (1) Are unlabeled data only useful when there are a large number of redundant features? (2) Are unlabeled data only useful when the classes are well separated?

We have recently presented an analysis of semi-supervised learning with maximum likelihood estimators that links performance degradation to incorrect modeling assumptions in a direct manner [6]. *Any* incorrect modeling assumption can lead to performance degradation in semi-supervised learning; there is no need to look for problems with "badly" separated classes, or problems with "bad" features. Even a small problem with two well-separated Gaussian features may suffer from unlabeled data [6].

When all modeling assumptions are correct, unlabeled data are always beneficial, no matter how separated are the classes and how poor are the features. However the picture changes when modeling assumptions are incorrect. The variance of maximum likelihood estimators still decreases with increasing number of labeled and unlabeled samples, but the level of bias introduced by incorrect modeling assumptions *depends on the ratio between the number of labeled and unlabeled samples*. The surprising fact is that bias depends not only on the differences between assumed and correct model, but also on the characteristics of the dataset — a fact that can be properly formalized using asymptotic analysis [6].

Proceedings of the ICML-2003 Workshop on The Continuum from Labeled to Unlabeled Data, Washington DC, 2003.

In practice we may find that some problems benefit from unlabeled data just because there is too little *labeled* data in the first place. This seems to be the case with text classification, where obviously incorrect modeling assumptions still display improvements with unlabeled data [9]. In such cases, the decrease in variance (caused by additional unlabeled samples) outweighs the increases in bias (caused by the same unlabeled samples).

3. Searching for correct models

Given that unlabeled data can be quite useful when the correct model is known, we should try to find the correct model whenever possible. Our experience suggests that model search is in fact *very* important in practice. We have consistently witnessed situations where labeled-unlabeled databases behave poorly under a fixed-structure classifier like the Naive Bayes, while model search schemes lead to excellent classifiers [4].

Here we would like to stress that semi-supervised learning requires a search for the *correct* model, not just a search within some arbitrary collection of models. In supervised learning it is possible to say that "if we are interested in the convergence of expected risks we should not resort to the convergence of probability measures" [7, p. 121], but such a statement should not be automatically extended to semi-supervised learning — we must consider the possibility that a more accurate probabilistic model will lead to significant gains from unlabeled data.

We observe that the discussion so far has focused on maximum likelihood estimators and generative classifiers. Could some other technique, perhaps a diagnostic one, offer a way out of model search? Empirical results suggest that performance degradation occurs in other semi-supervised paradigms as well. For example, Zhang and Oles [12] demonstrated that the use of unlabeled data with transductive SVM can cause degradation of performance when unlabeled data are added. Ghani [8] described experiments where the same phenomenon occurred with co-training. While no in-depth analysis has been made as to the cause of degradation in these cases, the following conjecture would unify these seemingly disparate results.

4. A conjecture

Here we wish to focus on methods that *effectively* handle unlabeled data — by that we mean methods that can benefit from unlabeled data at least when modeling assumptions are correct. We do not require that any such method knows the correct model; we just require that, *if* a method is applied to a problem that matches its assumptions, *then* the method successfully processes unlabeled data.

We propose the following informal conjecture: any method that effectively handles unlabeled data must either be informed of the correct model or must search for the correct model, in order to process arbitrary amounts of labeled and unlabeled data.

Note that a method may be quite succesful in benefiting from unlabeled data in situations where labeled data is extremely scarce, just because there is no information to extract from the labeled samples. For this reason, we require from any effective method that it must be ready to process small and large databases of labeled and unlabeled samples. *In such circumstances, we conjecture that model search must be conducted.*

As model search is a complex undertaking that requires significant effort, we would be happy to be proven wrong in our conjecture: it would be enough for someone to present a universally robust classifier — that is, a classifier that would always benefit from unlabeled data, even in cases where modeling assumptions are incorrect. Such a universally robust classifier would be a major accomplishment.

References

[1] S. Baluja. Probabilistic modeling for face orientation discrimination: Learning from labeled and unlabeled data. *Neural Information Processing Systems*, 1998.

[2] R. Bruce. Semi-supervised learning using prior probabilities and EM. *IJCAI Workshop on Text Learning*, 2001.

[3] V. Castelli. *The relative value of labeled and unlabeled samples in pattern recognition*. PhD thesis, Stanford, 1994.

[4] I. Cohen, N. Sebe, F.G. Cozman, M.C. Cirelo, and T.S. Huang. Learning Bayesian network classifiers for facial expression recognition using both labeled and unlabeled data. *CVPR*, 2003.

[5] F.G. Cozman and I. Cohen. Unlabeled data can degrade classification performance of generative classifiers. *FLAIRS*, 2002.

[6] F.G. Cozman and I. Cohen. The effect of modeling errors in semi-supervised learning of mixture models: How unlabeled data can degrade performance of generative classifiers. Report, http://www.poli.usp.br/p/fabio.cozman/Publications/ lul.ps.gz, 2003.

[7] R. Herbrich. Learning kernel classifiers: theory and algorithms. MIT Press, 2002.

[8] R. Ghani. Combining labeled and unlabeled data for multiclass text categorization. In *International Conference on Machine learning (ICML)*, 2002.

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

[10] J. Ratsaby and S.S. Venkatesh. Learning from a mixture of labeled and unlabeled examples with parametric side information. In *COLT*, pages 412–417, 1995.

[11] B. Shahshahani and D. Landgrebe. Effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. *IEEE Transactions on Geoscience and Remote Sensing*, 32(5):1087–1095, 1994.

[12] T. Zhang and F. Oles. A probability analysis on the value of unlabeled data for classification problems. In *ICML*, 2000.