Unlabeled Data Can Degrade Classification Performance of Generative Classifiers

Fabio G. Cozman[*]
Escola Politécnica, Universidade de São Paulo
Av. Prof. Mello Moraes, 2231 - 05508-900
São Paulo, SP - Brazil

Ira Cohen[†]
Hewlett-Packard Laboratories
1501 Page Mill Road
Palo Alto, CA 94304

Abstract

This paper analyzes the effect of unlabeled training data in generative classifiers. We are interested in classification performance when unlabeled data are added to an existing pool of labeled data. We show that unlabeled data can *degrade* the performance of a classifier when there are discrepancies between modeling assumptions used to build the classifier and the actual model that generates the data; our analysis of this situation explains several seemingly disparate results in the literature.

Introduction

The purpose of this paper is to discuss the performance of generative classifiers that are built with labeled and unlabeled records. Such classifiers have received attention in the machine learning literature due to their potential in reducing the need for expensive labeled data (Nigam *et al.* 2000; Seeger 2001). Applications such as web search, text classification, genetic research and machine vision are examples where we can find an abundance of cheap unlabeled data in addition to a pool of more expensive labeled data.

We show that there are cases where unlabeled data can *degrade* the performance of a classifier. We show that such a degradation can happen in common classification problems, and it is not a consequence of numerical instabilities, nor of outliers or other serious differences between assumed and actual models for data. Even minor modeling inaccuracies can lead to degradation from unlabeled data. We present an analysis of the labeled-unlabeled data problem, and demonstrate how unlabeled data can sometimes improve and sometimes degrade classification performance. Our analysis clarifies several seemingly disparate results that have been reported in the literature, and also explains existing but unpublished experiments in the field. The results offer insights on how to handle and use unlabeled data for classification and learning.

[*]This work was conducted while the first author was with the Internet Systems and Storage Laboratory, Hewlett-Packard Laboratories Palo Alto.

[†]Mailing address: The Beckman Institute, 405 N. Mathews Ave., Urbana, IL 61801.

Labeled-unlabeled data

The goal here is to label an incoming vector of *features* $\mathbf{X}$. Each instantiation of $\mathbf{X}$ is a *record*. We assume that there exists a *class variable* C; the values of C are the *labels*. We want to build *classifiers* that receive a record $\mathbf{x}$ and generate a label $\hat{c}(\mathbf{x})$ (notation follows (Friedman 1997)). Classifiers are built from a combination of existing labeled and unlabeled records.

If we knew exactly the joint distribution $p(C, \mathbf{X})$, we could design the optimal *classification rule* to label an incoming record $\mathbf{x}$ (Friedman 1997). Instead of storing the whole joint distribution $p(C, \mathbf{X})$, we could simply store the posterior distribution $p(C|\mathbf{X})$. This strategy is usually termed a *diagnostic* one (for example, diagnostic procedures are often used to "train" neural networks). In a statistical setting, diagnostic procedures may be cumbersome as they require a great number of parameters — essentially the same number of probability values as required to specify the joint distribution $p(C, \mathbf{X})$. An alternative strategy is to store the class distribution $p(C)$ and the conditional distributions $p(\mathbf{X}|C)$ and then, as we observe $\mathbf{x}$, compute $p(C|\mathbf{X} = \mathbf{x})$ using Bayes rule. This strategy is usually called *generative*. An advantage of generative methods is that unlabeled data do relate to some portions of the model (namely, the *marginal* distribution $p(\mathbf{X})$). If instead we focus solely on $p(C|\mathbf{X})$, there is no obvious and principled way to handle unlabeled data (Cohen, Cozman, & Bronstein 2002; Seeger 2001; Zhang & Oles 2000). For this reason, we employ generative schemes in this paper, and leave other approaches for future work.

So, we are interested in estimation of $p(C)$ and $p(\mathbf{X}|C)$. Note that it is possible for a classifier to have small estimation error and large classification error and vice-versa (Friedman 1997).

To build a classifier, we normally adopt a set of *modeling assumptions*. For example, we can assume a fixed number of labels. Or we assume a set of independence relations among variables; we call the set assumptions concerning independence relations the *structure* of a classifier. Once we fix our modeling assumptions, we estimate the parameters of the classifier. We assume here that all variables (class and features) have a known and fixed number of values, and that the structure that generates the data is fixed but not known. When the assumed structure matches the structure that gen-

erates the data, we say the structure is "correct."

The labeled-unlabeled data problem is a combination of both *supervised* and *unsupervised* problems (Duda & Hart 1973). Suppose that we have a classifier with modeling assumptions that exactly match the model generating data. Early work has proved that unlabeled data can lead to improved maximum likelihood estimates even in finite sample cases (Castelli & Cover 1996). (Shahshahani & Landgrebe 1994) emphasize the variance reduction caused by unlabeled data under the assumption that bias is zero; their conclusion is that unlabeled data must help classification (similar conclusion in (Zhang & Oles 2000)). In general, unlabeled data can help in providing information for the marginal distribution $p(\mathbf{X})$ (Cohen, Cozman, & Bronstein 2002). Castelli and Cover have investigated the value of unlabeled data in an asymptotic sense, with the assumption that the number of unlabeled records goes to infinity, and do so faster than the number of labeled records (Castelli 1994; Castelli & Cover 1995; 1996). They prove that, assuming modeling assumptions are correct, for the classifier, classification error decreases exponentially with the number of labeled records, and linearly with the number of unlabeled records[1] (similar analysis in (Ratsaby & Venkatesh 1995)). The message of previous work is that unlabeled data must help as long as modeling assumptions are correct.

On top of the theoretical work just described, several empirical investigations have suggested that unlabeled training data do improve classification performance. (Shahshahani & Landgrebe 1994) describe classification improvements with spectral data; Mitchell and co-workers report a number of approaches to extract valuable information from unlabeled data, from variations of maximum likelihood estimation (Nigam *et al.* 2000) to co-training algorithms (Mitchell 1999). Other publications report on EM-like algorithms (Baluja 1998; Bruce 2001; Miller & Uyar 1996) and co-training approaches (Collins & Singer 2000; Comité *et al.* 1999; Goldman & Zhou 2000). There have also been workshops on the labeled-unlabeled data problem (at NIPS1998, NIPS1999, NIPS2000 and IJCAI2001).

Overall, these publications and meetings advance an optimistic view of the labeled-unlabeled data problem, where unlabeled data can be profitably used whenever available. A more detailed analysis of current results does reveal some puzzling aspects of unlabeled data. In fact, the workshop held at IJCAI2001 witnessed a great deal of discussion on whether unlabeled data are really useful.[2]

Three results are particularly interesting:

1. Shahshahani and Landgrebe describe experiments that demonstrate how unlabeled data can help mitigate the "Hughes phenomenon" (degradation of performance when features are added), but they also report situations where unlabeled data degrade performance. They attribute such cases to deviations from modeling assumptions; for example, "outliers, . . . , and samples of unknown classes" — they even suggest that unlabeled records should be used with care, and only when the labeled data alone produce a poor classifier.
2. (Baluja 1998) used *Naive Bayes* (Friedman 1997) and *TAN* classifiers (Friedman, Geiger, & Goldszmidt 1997) to obtain excellent classification results, but there were cases where unlabeled data degraded performance.
3. In work aimed at classification of documents, (Nigam *et al.* 2000) used Naive Bayes classifiers with fixed structure and a large number of features. Nigam et al actually discuss situations where unlabeled data degrade performance, and propose several techniques to reduce the observed degradation. Nigam et al do not attempt to completely explain the reasons for degradation, but suggest that the problem might have been a mismatch between the natural clusters in feature space and the actual labels.

The present paper can be understood as a natural sequence on Nigam et al investigation, where we verify and explain the conditions that lead to degradation with unlabeled data. In fact, intrigued by these existing results, we conducted a series of experiments aimed at understading the value of unlabeled data (Cohen, Cozman, & Bronstein 2002). In short, the experiments do indicate that unlabeled data can have a deleterious effect in some situations. Consider Figure 1, which shows two typical results. Here we estimated the parameters of Naive Bayes classifiers with 10 features using the EM algorithm (Dempster, Laird, & Rubin 1977). Figure 1 shows classification performance when the underlying model actually has a Naive Bayes structure (left), and when the underlying model follows a TAN model (right). The result is clear: when we estimate a Naive Bayes classifier with data from a Naive Bayes model, more unlabeled data help; when we estimate a Naive Bayes classifier with data that do not come from a corresponding structure, more unlabeled data can degrade performance.

The previous discussion raises some questions: Can unlabeled data actually degrade performance, and if so, how, and why?

The effect of unlabeled data

In this section we discuss the effect of unlabeled data on classification error. To visualize the effect of unlabeled data, we propose a new strategy for graphing performance in the labeled-unlabeled data problem. Instead of fixing the number of labeled records and varying the number of unlabeled records, we propose to fix the *percentage* of unlabeled records among all training records. We then plot classification error against the number of training records. Call such a graph a *LU-graph*. We introduce LU-graphs with an example.

Consider a situation where we have a class variable C with labels c_0 and c_1, and probability $p(c_0) = 0.4017$. We also have two real-valued features X_1 and X_2 with distributions:

$$p(X_1|c_0) = N(2,1), \quad p(X_1|c_1) = N(3,1),$$

[1] Castelli and Cover assume identifiability, a property that may fail when features are discrete, as in many machine learning applications (Duda & Hart 1973). Lack of identifiability does not seem to be a crucial matter in the labeled-unlabeled problem, as we made extensive tests with discrete models and observed behavior consistent with Gaussian (identifiable) models.

[2] This fact was communicated to us by George Forman.

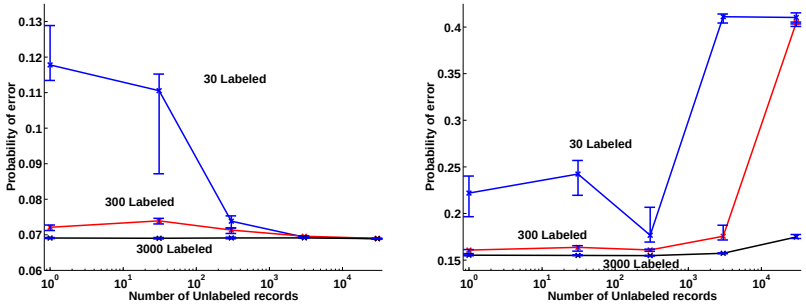

Figure 1: Naive Bayes classifier from data generated from a Naive Bayes model (left) and a TAN model (right). Each point summarizes 10 runs of each classifier on testing data; bars cover 30 to 70 percentiles.

$$p(X_2|c_0, x_1) = N(2, 1), \quad p(X_2|c_1, x_1) = N(1 + 2x_1, 1),$$

where $N(\mu, \sigma^2)$ denotes a Gaussian distribution with mean μ and variance σ^2. The problem is identifiable, and there is dependency between X_2 and X_1 (X_2 depends on X_1 when $C = c_1$). Suppose we build a Naive Bayes classifier for this problem. Figure 2 shows LU-graphs for 0% unlabeled records, 50% unlabeled records and 99% unlabeled records. The LU-graphs for 50% and 99% unlabeled data have an interesting property: their asymptotes do not converge to the same value, and they are both different from the asymptote for labeled data. Suppose then that we started with 50 labeled records as our training data. Our classification error would be about 7.8%, as we can see in the LU-graph for 0% unlabeled data. Suppose we added 50 labeled records, and we reduced classification error to about 7.2%. Now suppose we added 100 *unlabeled* records. We would move from the 0% LU-graph to the 50% LU-graph. Classification error would increase to 8.2%! And if we then added 9800 unlabeled records, we would move to the 99% LU-graph, with classification error about 16.5% — more than twice the error we had with just 50 labeled records.

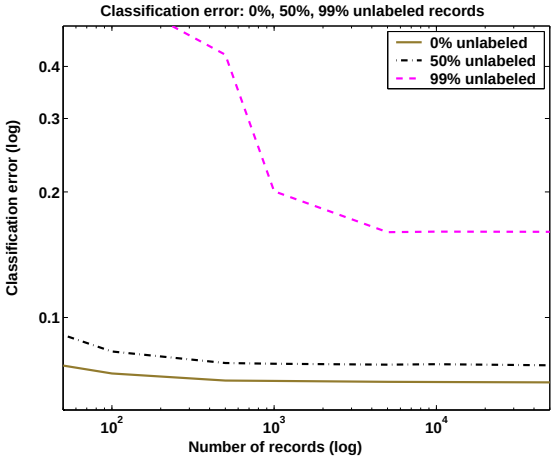

Figure 2: LU-graphs for the example with two Gaussian features. Each point in each graph is the average of 100 trials; classification error was obtained by testing in 10000 labeled records drawn from the correct model.

The fact that classification error has different asymptotes, for different levels of unlabeled data, can lead to a degradation of performance from the addition of unlabeled data. By moving from one LU-graph to another, we can either see an increase or a decrease on classification error, depending on the slopes of the particular LU-graphs.

It should be noted that in difficult classification problems, where LU-graphs decrease very slowly, unlabeled data should improve classification performance (unless we have large amounts of labeled data available). Problems with a large number of features and parameters should require more training data, so we can expect that such problems benefit more consistently from unlabeled data. Examples discussed by (Nigam *et al.* 2000) seem to fit this description exactly — while they suggest that adding features can worsen the effect of unlabeled data, the opposite should be expected. This observation also agrees with the empirical findings of (Shahshahani & Landgrebe 1994), where unlabeled data are useful as more and more features are used in classifiers.

To understand why and when do unlabeled data produce asymptotic differences between LU-graphs, we must visualize the geometry of estimation and adopt some assumptions regarding the problem. We simplify the analysis by concentrating on the two extreme LU-graphs, for fully labeled data and for fully unlabeled data — we should expect that any asymptotic gap of performance between these two graphs will be filled by a continuum of LU-graphs (just as the middle graph in Figure 2).

In general, if we do not have the correct modeling assumptions for a classifier, we can obtain a model that only approximates $p(C, \mathbf{X})$, regardless of the estimation procedure we employ. We can imagine a set of probability distributions $K(C, \mathbf{X})$ that represents the models compatible with modeling assumptions; we are assuming that $p(C, \mathbf{X}) \notin K(C, \mathbf{X})$. It is easy to imagine such a set of distributions as a polytope in a high-dimensional space, even though modeling assumptions may induce a more complex set. Figure 3 shows the set $K(C, \mathbf{X})$ and the "correct" distribution $p(C, \mathbf{X})$ outside of $K(C, \mathbf{X})$. We can also imagine a different set, $K(\mathbf{X})$, obtained by pointwise marginalization of $K(C, \mathbf{X})$. Figure 3 also shows this other set and the distribution $p(\mathbf{X})$ obtained by marginalization of $p(C, \mathbf{X})$.

Our first assumption is that the closest distribution to

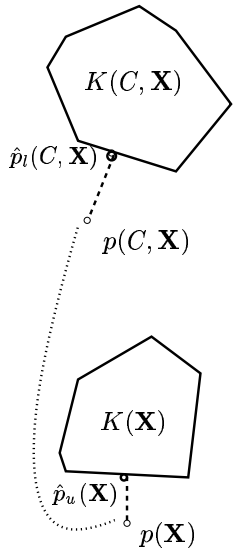

Figure 3: Sets of distributions induced by modeling assumptions, and distributions generated by estimation.

$p(C, \mathbf{X})$ in $K(C, \mathbf{X})$ (which we denote by $\hat{p}_l(C, \mathbf{X})$) does not correspond to the closest distribution to $p(\mathbf{X})$ in $K(\mathbf{X})$ (which we denote by $\hat{p}_u(\mathbf{X})$). Here "closeness" between distributions is induced by the estimation procedure, and it need not be a true norm. To illustrate this assumption, take maximum likelihood estimation in discrete models. Consider first the maximum likelihood estimate for $p(C, \mathbf{X})$ in $K(C, \mathbf{X})$. As the number of training records grows without bound, the empirical distribution converges to $p(C, \mathbf{X})$, so asymptotically we will choose the distribution in $K(C, \mathbf{X})$ that is closest to $p(C, \mathbf{X})$ with respect to the Kullback-Leibler divergence (Friedman, Geiger, & Goldszmidt 1997).

Suppose now we compute the maximum likelihood estimate for $p(\mathbf{X})$ subject to the choices in $K(\mathbf{X})$, the maximum likelihood solution for unlabeled data. We obtain the closest distribution to $p(\mathbf{X})$ in $K(\mathbf{X})$ with respect to Kullback-Leibler divergence. Note that $\hat{p}_u(\mathbf{X})$ induces estimates $\hat{p}_u(C, \mathbf{X})$. As Kullback-Leibler divergence and marginalization do not commute, $\hat{p}_l(C, \mathbf{X})$ and $\hat{p}_u(C, \mathbf{X})$ need not be equal — and we must focus exactly on the situation where they are different.[3]

An additional argument is needed to understand the effect of unlabeled data: We must appreciate the difference between classification and estimation. Even though adding more data (labeled and unlabeled) leads to better overall estimation (with respect to various global measures such as likelihood, squared-error, variance, Fisher information), the improvement may be uneven amongst the estimated parameters. Note that for classification, only $p(C|\mathbf{X})$ matters (Friedman 1997); if the bias in $\hat{p}_u(C|\mathbf{X})$ is larger than the bias in $\hat{p}_l(C|\mathbf{X})$, the asymptotic classification performance for unlabeled data is smaller than for labeled data. When this performance gap is present, then unlabeled data can degrade performance and the LU-graphs can be used to capture this phenomenon. Basically, the fact that *estimation error* is the guiding factor in building a classifier leads us to use estimates that are not optimal with respect to *classification error*.[4]

The preceeding discussion indicates that missing labels are different from missing feature values. Both forms of missing data degrade estimation performance, but unlabeled data also affects classification performance directly by introducing bias in the critical parameters $p(C|\mathbf{X})$. This insight clarifies questions on missing/unlabeled data raised by (Seeger 2001).

Conclusion

The message of this paper is that unlabeled training data can degrade classification performance if we have modeling assumptions for a classifier. There have been reports that unlabeled data can produce such a degradation, but explanations offered so far suggest that degradation should occur in somewhat extreme circunstances. The main point of this paper is that the type of degradation produced by unlabeled data can occur under common assumptions and can be explained by fundamental differences between classification and estimation errors. We propose LU-graphs as an excellent visualization of this phenomenon. An important point for future investigation is the asymptotic classification performance for various percentages of unlabeled data. Another point is the search for other possible sources of performance degradation, particularly when there are severe mismatches between actual and assumed models.

It certainly seems that some creativity must be exercised when dealing with unlabeled data. As discussed in the literature (Seeger 2001), currently there is no coherent strategy for handling unlabeled data with diagnostic classifiers, and generative classifiers are likely to suffer from the effects described in this paper. More general, or simply different, approaches could be welcome (Jaakkola, Meila, & Jebara 1999). Future work should investigate whether unlabeled data can degrade performance in different classification approaches, such as decision trees and co-training. Regardless of the approaches that are used, unlabeled data are affected by modeling assumptions, and we can use unlabeled data to help our search for a correct modeling assumptions. The present paper should be helpful as a first step in the understanding of unlabeled data and their peculiarities in machine learning.

Acknowledgements

We thank Alex Bronstein and Marsha Duro for proposing the research on labeled-unlabeled data and for many suggestions and comments during the course of the work; their

[3] Note that when we have correct modeling assumptions, $p(C, \mathbf{X}) \in K(C, \mathbf{X})$, and then both estimates must be equal assuming identifiability.

[4] This conclusion is not restricted to maximum likelihood estimation, nor does it depend on identifiability. We omit examples with least-squares estimation and identifiable models due to lack of space.

help was critical to the work described here. We thank Vittorio Castelli for sending us a copy of his PhD dissertation, Charles Elkan for sending us his BNB software, and George Forman for telling us about the IJCAI workshop on unlabeled data. We thank Kevin Murphy for the freely available BNT system, which we used to generate examples and data. We coded our own Naive Bayes and TAN classifiers in the Java language, using the libraries of the JavaBayes system (freely available at http://www.cs.cmu.edu/~javabayes). Marina Meila read a preliminary version and sent useful comments.

References

Baluja, S. 1998. Probabilistic modeling for face orientation discrimination: Learning from labeled and unlabeled data. In *Neural and Information Processing Systems (NIPS)*.

Bruce, R. 2001. Semi-supervised learning using prior probabilities and EM. In *IJCAI-01 Workshop on Text Learning: Beyond Supervision*.

Castelli, V., and Cover, T. M. 1995. On the exponential value of labeled samples. *Pattern Recognition Letters* 16:105–111.

Castelli, V., and Cover, T. M. 1996. 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.

Castelli, V. 1994. *The Relative Value of Labeled and Unlabeled Samples in Pattern Recognition*. Ph.D. Dissertation, Stanford University.

Cohen, I.; Cozman, F. G.; and Bronstein, A. 2002. The effect of unlabeled data on generative classifiers, with application to model selection. Technical report, HP labs.

Collins, M., and Singer, Y. 2000. Unupervised models for named entity classification. In *Proc. 17th International Conf. on Machine Learning*, 327–334. Morgan Kaufmann, San Francisco, CA.

Comité, F. D.; Denis, F.; Gilleron, R.; and Letouzey, F. 1999. Positive and unlabeled examples help learning. In Watanabe, O., and Yokomori, T., eds., *Proc. of 10th International Conference on Algorithmic Learning Theory*, 219–230. Springer-Verlag.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal Royal Statistical Society B* 44:1–38.

Duda, R. O., and Hart, P. E. 1973. *Pattern Classification and Scene Analysis*. New York: John Wiley and Sons.

Friedman, N.; Geiger, D.; and Goldszmidt, M. 1997. Bayesian network classifiers. *Machine Learning* 29:131–163.

Friedman, J. H. 1997. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery* 1(1):55–77.

Goldman, S., and Zhou, Y. 2000. Enhancing supervised learning with unlabeled data. In *International Joint Conference on Machine Learning*.

Jaakkola, T. S.; Meila, M.; and Jebara, T. 1999. Maximum entropy discrimination. *Neural Information Processing Systems 12*.

Miller, D. J., and Uyar, H. S. 1996. A mixture of experts classifier with learning based on both labelled and unlabelled data. In *Advances in Neural Information Processing Systems*. 571–577.

Mitchell, T. 1999. The role of unlabeled data in supervised learning. In *Proc. of the Sixth International Colloquium on Cognitive Science*.

Nigam, K.; McCallum, A. K.; Thrun, S.; and Mitchell, T. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning* 39:103–144.

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

Seeger, M. 2001. Learning with labeled and unlabeled data. Technical report, Institute for Adaptive and Neural Computation, University of Edinburgh, Edinburgh, United Kingdom.

Shahshahani, B. M., and Landgrebe, D. A. 1994. The 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.

Zhang, T., and Oles, F. 2000. A probability analysis on the value of unlabeled data for classification problems. In *International Joint Conference on Machine Learning*, 1191–1198.