Splitting the Unsupervised and Supervised Components of Semi-Supervised Learning

Clayton Silva Oliveira CLAYTON.OLIVEIRA@POLI.USP.BR
Fabio Gagliardi Cozman FGCOZMAN@USP.BR
Escola Politécnica, University of Sao Paulo, Cidade Universitaria, Sao Paulo, SP - Brazil

Ira Cohen IRA.COHEN@HP.COM
Hewlett-Packard Laboratories, 1501 Page Mill Road, Palo Alto, CA 94304

Abstract

In this paper we investigate techniques for semi-supervised learning that split their unsupervised and supervised components — that is, an initial unsupervised phase is followed by a supervised learning phase. We first analyze the relative value of labeled and unlabeled data. We then present methods that perform "split" semi-supervised learning and show promising empirical results.

1. Introduction

Many techniques for semi-supervised learning are based on principles that apply without distinction to labeled and unlabeled data (maximum likelihood, posterior expected loss, maximum entropy). Often the idea is to process both kinds of data either together or in close collaboration, under the guidance of one general principle. However one can conceive a different strategy, where labeled and unlabeled data contribute in distinct ways to the learning procedure. An example of this "split" strategy is the recent work on learning Riemannian manifolds, where manifolds are detected only with the unlabeled data, and then used in a transductive method in a second step (Belkin & Niyogi, 2004). As stated by Ando and Zhang (2005), "the basic idea is to learn good functional structures using the unlabeled data" and to employ the resulting structures in a later stage.

At first it might seem that labeled and unlabeled data should always be given the same status: in the absence of modeling errors, both kinds of data are valuable in reducing classification error (Castelli & Cover, 1996). However this intuitive argument breaks in the presence of modeling errors, as in this case unlabeled data may have a deleterious effect on performance — Section 2 presents a brief analysis on the relative value of labeled and unlabeled data in generative classifiers. Thus in practice labeled and unlabeled data have different effects, and it is advisable to consider strategies that do not treat them equally.

The goal of this paper is to test the possibly simplest way to "split" labeled and unlabeled learning components. We wish to explore how far one can go by processing attributes with a well-known unlabeled method (PCA/ICA) and using the transformed attributes in a supervised fashion (here by SVMs). In Sections 3 and 4 we present basic concepts and experiments on this strategy.

2. The value of labeled and unlabeled data

The goal in this paper is to classify an incoming vector of observables $\mathbf{X}$. Each instantiation of $\mathbf{X}$ is a *sample*. There exists a *class variable* C that takes values in a set of *labels*. To simplify the discussion, we assume that C is a binary variable with values c' and c''. We want to build *classifiers* that receive a sample $\mathbf{x}$ and output a value for C. We assume 0-1 loss, hence our objective is to minimize the probability of classification errors. If we knew exactly the joint distribution $P(C, \mathbf{X})$, the optimal rule would be to select the label with highest posterior probability (Devroye et al., 1996).

A classifier is to be built using samples in a database. The samples in the database are either *labeled* or *unlabeled*; we assume that a sample is unlabeled with probability $(1 - \lambda)$. We also assume that the same distribution $P(\mathbf{X}|C)$ generates both kinds of samples.

Appearing in *Proceedings of the 22^{st} International Conference on Machine Learning*, Bonn, Germany, 2005.

When a sample is unlabeled, its distribution is a mixture $\eta p(\mathbf{X}|c') + (1-\eta)p(\mathbf{X}|c'')$, where $\eta = p(c')$. We assume that such mixtures are identifiable (Redner & Walker, 1984).

Consider that a (parametric) generative model $P(C, \mathbf{X}|\theta)$ is adopted as a representation for the joint probability $P(C, \mathbf{X})$. Two guidelines often used to generate estimates $\hat{\theta}$ are *maximum likelihood* and *maximization of posterior loss* (DeGroot, 1970). In recent years a number of authors has explored generative models and maximum likelihood (or variants thereof) in semi-supervised learning, with promising results (Baluja, 1998; Bruce, 2001; Miller & Uyar, 1996; Nigam et al., 2000). Theoretical results show that labeled and unlabeled data contribute to a reduction in risk, even though they may do so at different rates, whenever modeling assumptions are *correct* — that is, when $P(C, \mathbf{X}|\theta)$ is equal to $P(C, \mathbf{X})$ for some θ (Castelli & Cover, 1995; Castelli & Cover, 1996; Ratsaby & Venkatesh, 1995).

However, these positive findings on generative classifiers have been contrasted with a number of examples where unlabeled data has led to some degradation in classification error (Grandvalet & Bengio, 2004; Cozman & Cohen, 2002; Shahshahani & Landgrebe, 1994). Here "degradation" means that the classification error obtained with labeled and unlabeled data is larger than the classification error obtained just with labeled data. Recently an asymptotic analysis of semi-supervised learning has focused on the effect of modeling errors on performance degradation in semi-supervised learning (Cozman et al., 2003). The analysis can be summarized as follows. In the presence of modeling errors, asymptotic estimates obtained by maximum likelihood are affected by the probability λ: estimates change as λ moves from the supervised situation ($\lambda = 1$) to the unsupervised limit ($\lambda \to 0$). Suppose one starts with labeled data and gradually adds unlabeled data. The effect is a gradual change in λ and a corresponding gradual change in estimates, from their supervised starting point in the direction of an unsupervised limit. This explains why in practice one may find that taking larger and larger amounts of unlabeled data changes not only the variance of estimates but also their average behavior. Now, is the supervised starting point better than the unsupervised limit? Intuitively one would expect labeled data to provide more guidance to a learning procedure, thus producing better asymptotic estimates than the unlabeled data. This rationale is discussed in more detail in Appendix A.

The possibility of degradation in generative classifiers suggests that we should investigate more the performance of diagnostic classifiers for semi-supervised learning. Nevertheless, as purely diagnostic classifiers are not affected by unlabeled data (Zhang & Oles, 2000), is not trivial to design the inclusion of unlabeled data into the process. There must be some "principle" connecting the probabilities over C and $\mathbf{X}$, and the probabilities over $\mathbf{X}$. There is a great diversity of successful solutions, ranging from maximum entropy solutions (Grandvalet & Bengio, 2004; Jaakkola et al., 1999) to transductive methods (Joachims, 1999).[1] Hence it is relatively hard to select the best diagnostic approach to unlabeled data, given the number of different strategies in the literature.

3. "Split" learning

The previous section presented some of the challenges in semi-supervised learning. On the one hand, it seems generally difficult to guarantee that generative classifiers will be immune to performance degradation. On the other hand, it is not easy to modify diagnostic classifiers in a coherent fashion so as to make them "sensitive" to unlabeled data.

Now, unlabeled data can be an extremely valuable source of insight on modeling assumptions. Unlabeled data may be used to verify modeling assumptions in generative classifiers (as proposed by Cohen et al. (2004)); or unlabeled data may be used to establish modeling assumptions prior to actual use of a learning method. Ando and Zhang (2005) refer to this latter strategy as "structural" learning, as it looks at structural aspects of the learning situation using the unlabeled data. A similar idea on manifolds is proposed by Belkin and Niyogi (2004). We consider in this paper a particular case where an initial unsupervised phase is followed by a fully supervised, diagnostic phase; we refer to this as "split" learning.

Our purpose is to explore the performance of a rather simple strategy that can be applied without any modification on existing methods. We consider that the first unsupervised phase focuses on the transformation of attributes using "classic" methods such as PCA (Hastie et al., 2003) and ICA (Hyvarinen, 1999), while the second supervised phase uses SVMs (Vapnik, 1998). Similar ideas have been applied previously by

[1] A few empirical results suggest that performance degradation may occur also in diagnostic paradigms, but no in-depth analysis has been conducted — for example, Zhang and Oles (2000) discuss performance degradation with transductive SVMs, while Ghani (2002) describes experiments where the same phenomenon occurred with co-training.

L'Heureux et al. (2004).

"Split" learning has a few advantages. A great deal of unlabeled data can be used in an exploratory first step, and then a classifier can be quickly learned with the available labeled samples. The supervised phase can use either a generative or diagnostic method without any difficulty. Conclusions obtained with unlabeled data can be transferred to a variety of supervised classifiers.

To a great extent, the goal of this paper is to test the *simplest* conceivable scheme for "split" learning. The resulting procedure may be criticized as too simplistic and prone to elementary mistakes (for example, PCA may collapse relevant clusters together). However, we note that a similar criticism can be leveled against the successful Naive Bayes classifier as adopting unrealistic assumptions. The key motivation here is to explore a method that can be quickly employed by any practitioner with existing tools at hand. The next section shows that the PCA/SVM combination is quite promising in practice.

4. Experiments

The idea here is to use PCA or ICA to transform the attributes in a compact and effective manner, and then to use SVMs in a supervised fashion with the transformed attributes. Both PCA and ICA try to use training data to find a matrix T that transforms this data to a different space where the features are *statistically independent*. Although PCA and ICA have the same objective, they differ in their assumptions: PCA assumes that mean and variance are sufficient statistics, thus in fact assuming Gaussianity; ICA does not assumes the data to be Gaussian and finds a linear transformation such that statistical independence over features is maximized. While PCA is well suited as a method to reduce the number of attributes in a problem, ICA is less suited for this task (Hyvarinen, 1999).

In our experiments we have used the PCA implementation in MATLAB R13 and the ICA implementation in `FastICA` (Hyvarinen & Oja, 1997). The SVM classifiers are learned by the `MATLAB Support Vector Machine Toolbox` (Cawley, 2000). We run experiments on the `20 Newsgroups` database (tokenized with the `Rainbow` package (McCallum, 1996)) and with three real databases from the UCI Database Repository (Blake & Merz, 1998): `Adult`, `Spam` and `Isolet` databases. In all experiments the unlabeled samples were chosen randomly.

We studied six different cases, combining SVM as classifier and PCA or ICA for feature transformations and dimensionality reduction. We have

- **PCA Lab**: We executed PCA *only* with the labeled data; then we selected a number of principal components (PCs), ranked by the respective eigenvalues, so the dimensionality of the labeled database was reduced. We then learned an SVM with the transformed labeled database.
- **PCA LUL**: The same procedure in the previous case, but using the partially labeled database to execute PCA. The labeled data, transformed by PCA, was used to learn a supervised SVM.
- **ICA Lab**: We executed ICA *only* with the labeled data. Then we selected a number of independent components (ICs), ranked by the respective values of the inverse of the absolute kurtosis[2], so the dimensionality of the labeled database was reduced. We then learned an SVM with the transformed labeled database.
- **ICA LUL**: The same procedures in the previous case, but using the partially labeled database to execute ICA. The labeled data, transformed by ICA, was used to learn a supervised SVM.
- **ICA-all Lab**: The same procedure in the "ICA Lab" case, but without dimensionality reduction in the labeled database.
- **ICA-all LUL**: The same procedures in the "ICA LUL" case, but without dimensionality reduction in the labeled database.

The SVMs were learned with RBF (radial basis function) kernels: $K(x, x') = e^{-\|x-x'\|^2/2\sigma^2}$.

Each point in the graphs presented later represents the average performance of SVMs by 10-fold cross validation. The dimension reduction techniques are separately applied within each fold of cross-validation. The vertical axis gives us the average *increase* of performance *over* the average performance of a supervised SVM, i.e., taking *only* the labeled data in their original feature space to train the classifier. For example, if we have 5% in the vertical axis, we just have the performance of the supervised SVM *plus* 5%. The horizontal axis contains the number of selected components to each point of the curves (not valid to the "ICA-all Lab" and "ICA-all LUL" cases).

[2]The kurtosis of a database is computed by $kurt(x) = E\{x^4\} - 3(E\{x^2\})^2$. If $p(x)$ is a Gaussian distribution, $kurt(x)$ is equal to 3.

We emphasize that the 0% in the vertical axis denotes the performance of an SVM learned with the few labeled samples selected for each experiment. We refer to these SVMs as "totally supervised" SVMs.

4.1. Results with the *20 Newsgroups* database

To perform experiments with the `20 Newsgroups` database, we considered the messages in the `rec.autos` and `rec.motorcycles` directories. We must classify a given message drawn from one of these directories as a message about `autos` or `motorcycles`. Using the `Rainbow` package, we tokenized these directories and limited the number of features to 1250.

We considered 40 labeled samples in all experiments with this database, and 100, 500 and 1000 unlabeled data. The measured average kurtosis of this database was (439.58 ± 544.72).

Figure 1 shows the results of "split" learning in the `20 Newsgroups` database. The inclusion of unlabeled data to perform PCA improved performance significantly. We could achieve up to 25% improvements over the "totally supervised" SVM, using just the 5 first PCs (less than 1% of the original number of features – 1250). Moreover, results seem to be similar to those obtained through the structural learning method by Ando and Zhang (2005).

Learning ICA from only the labeled data and not performing dimensional reduction (curve "ICA-all Lab" from Figure 1) was successful to achieve a better performance SVM, although in this case we had more features than PCA case. Anyway, PCA was better to incorporate unlabeled data to produce indirectly semi-supervised SVMs, moreover reducing drastically the dimension of the labeled data.

4.2. Results with the *Spam* database

The `Spam` database contains 56 features. We considered only 22 labeled data in all experiment, and took 100, 500 and 3228 unlabeled samples. The measured average kurtosis of this database was (218.93 ± 326.31).

Figure 2 shows the results of "split" learning with the `Spam` database. Performing PCA and dimensionality reduction prior to the learning of the supervised SVM increased the average performance of this classifier. Adding unlabeled data to perform PCA enhanced even more the performance of the SVM. For example, the inclusion of 100 unlabeled instances to execute PCA (top-left graph in Figure 2) increased the performance of the SVM up to 15% with only the first PC, when compared to the "totally supervised" SVM. While PCA was quite successful, ICA did not present

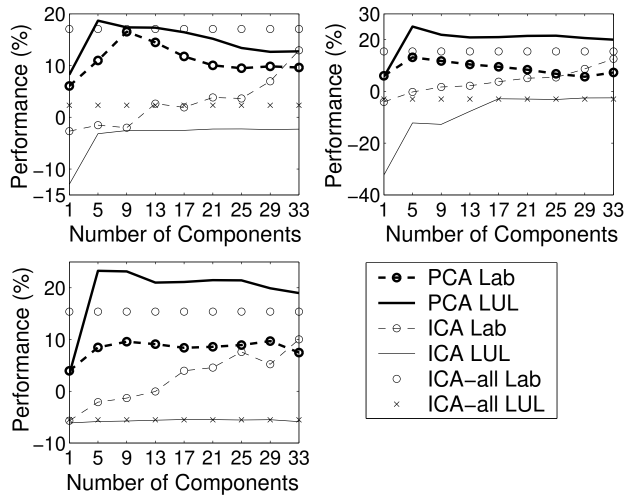

Figure 1. Results obtained with the `20 Newsgroups` database. *Top-left graph*: 40 labeled instances plus 100 unlabeled ones. The performance of the "totally supervised" SVM was $(52.82 \pm 3.30)\%$. *Right graph*: 40 labeled instances plus 500 unlabeled ones. The performance of the "totally supervised" SVM was $(52.57 \pm 2.60)\%$. *Bottom-left graph*: 40 labeled instances plus 1000 unlabeled ones. The performance of the "totally supervised" SVM was $(55.27 \pm 3.22)\%$.

positive results.

4.3. Results with the *Adult* database

The `Adult` database has 14 features. We used only 20 labeled data in all experiment, and 100, 1000 and 10000 unlabeled samples. The measured average kurtosis of this database was (9.22 ± 19.85).

Figure 3 shows the results with the `Adult` database. Including unlabeled data to perform PCA led to an increase of performance when considering just the first PCs, although in some cases the effect of the additional unlabeled data was deleterious to the performance of the SVM, in comparison to the case when performing PCA only with the labeled data. The inclusion of 10000 unlabeled instances to execute PCA (bottom-left graph in Figure 3) increased the SVM performance up to almost 10% with only the first PC. ICA was again unsuccesful (even though including unlabeled data to execute ICA improved the SVM performance when compared to the case we used only the labeled data to perform ICA – curves "ICA Lab" vs. "ICA LUL").

4.4. Results with the *Isolet* database

The `Isolet` database contains 617 features. We fixed 20 labeled samples in all experiments, and 100, 500

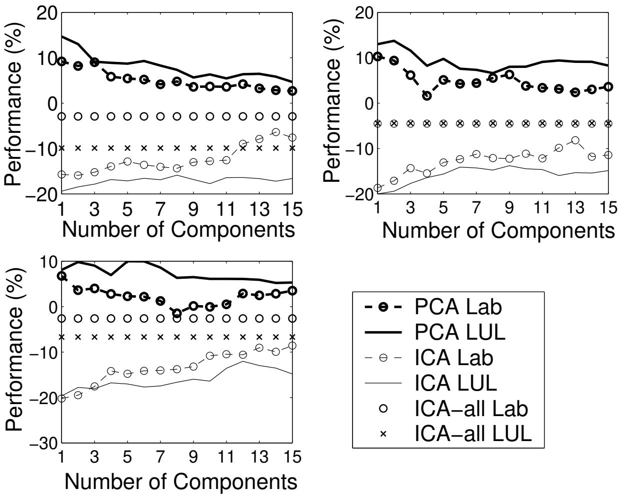

Figure 2. Results obtained with the `Spam` database. *Top-left graph*: 22 labeled instances plus 100 unlabeled ones. The performance of the "totally supervised" SVM was $(69.79 \pm 6.15)\%$. *Right graph*: 22 labeled instances plus 500 unlabeled ones. The performance of the "totally supervised" SVM was $(70.22 \pm 6.36)\%$. *Bottom-left graph*: 22 labeled instances plus 3228 unlabeled ones. The performance of the "totally supervised" SVM was $(71.32 \pm 5.69)\%$.

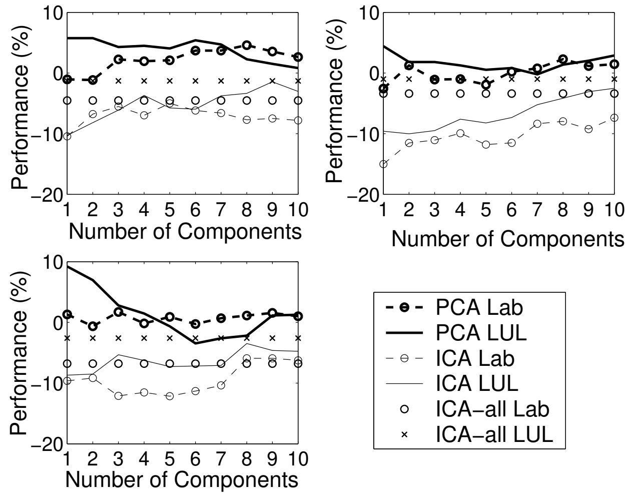

Figure 3. Results obtained with the `Adult` database. *Top-left graph*: 20 labeled instances plus 100 unlabeled ones. The performance of the "totally supervised" SVM was $(62.41 \pm 6.24)\%$. *Right graph*: 20 labeled instances plus 1000 unlabeled ones. The performance of the "totally supervised" SVM was $(64.17 \pm 4.81)\%$. *Bottom-left graph*: 20 labeled instances plus 10000 unlabeled ones. The performance of the "totally supervised" SVM was $(62.72 \pm 5.62)\%$.

and 1000 unlabeled samples. The measured average kurtosis of this database was (4.43 ± 15.16).

Figure 4 shows the results we obtained using indirectly semi-supervised SVM and the `Isolet` database. The inclusion of unlabeled data to execute PCA seems not to cause any effect in the learning of the SVM, as we observe comparing "PCA Lab" and "PCA LUL" curves. ICA was more efficient than PCA only when the number of ICs was higher than 11, without including any unlabeled data ("ICA Lab" case).

5. Discussion

The vagaries of semi-supervised learning suggest that a valid strategy is to always start with a supervised classifier (learned with labeled data only). This "baseline" classifier can then be compared to other semi-supervised classifiers. Whenever modeling assumptions seem inaccurate, the use of unlabeled data as an exploratory tool is a profitable decision. "Split" learning represents an extreme alternative where the unlabeled data is employed as a modeling tool while the labeled data is used for supervised learning.

Our experiments suggest that, despite its startling simplicity, "split" learning with PCA+SVM is quite effective; in fact it seems to be as effective as much more complex proposals in the literature (e.g., structural learning (Ando & Zhang, 2005)). Experiments show that PCA is quite effective in "rewriting" attributes in a compact and profitable manner. In those practical problems where dimensional reduction is desirable (for instance, when the number of attributes is high), the use of our simple "split" learning method is natural and promising. Once the attributes are rewritten in a compact manner, thus conveying the content of the unlabeled data, the transformed attributes can be quickly used to train other supervised classifiers with few low dimensional labeled samples, without returning to the unlabeled data. Experiments such as the `20 Newsgroups` database show that in some situations we may be able to employ less than 1% of the original number of attributes and still improve performance. We emphasize that in this and in other experiments the "quality" of the PCA reduction is greatly improved by the presence of unlabeled data.

An interesting observation is that databases that comply less with the Gaussian assumptions of PCA (measured by the average kurtosis) tend to benefit more significantly from unlabeled data. When the database is close to the Gaussianity assumption, a few labeled samples are enough to produce a satisfactory tranformation for attributes. When the database is not Gaussian at all, PCA needs more samples to work with — exactly the samples provided by the unlabeled data.

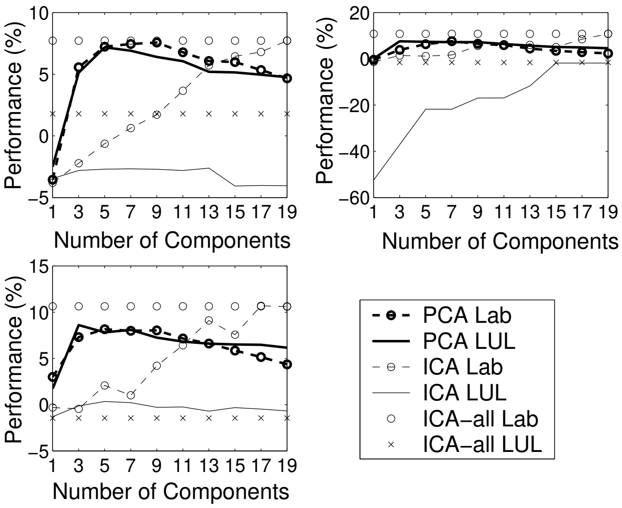

Figure 4. Results obtained with the `Isolet` database. *Top-left graph*: 20 labeled instances plus 100 unlabeled ones. The performance of the "totally supervised" SVM was $(54.02 \pm 6.61)\%$. *Right graph*: 20 labeled instances plus 500 unlabeled ones. The performance of the "totally supervised" SVM was $(52.43 \pm 4.86)\%$. *Bottom-left graph*: 20 labeled instances plus 1000 unlabeled ones. The performance of the "totally supervised" SVM was $(50.06 \pm 4.25)\%$.

We note that ICA was not effective in improving performance. Future work should investigate why PCA and ICA are so different in the present context.

We also leave for future work the search for more sophisticated (but hopefully still simple) methods that transform attributes taking into consideration the second supervised step.

Acknowledgments

The first author is supported by FAPESP (www.fapesp.br), under grant number 03/09653-5. The second author is partially supported by a scholarship from CNPq (www.cnpq.br). This work has received generous support from HP Brazil R&D.

A. Labeled and unlabeled data in generative classifiers

In this appendix we return to a question stated in Section 2: in generative classifiers, is the supervised starting point better than the unsupervised limit of performance? One would expect labeled data to have more useful content than unlabeled data. But even a simple example shows that the matter is not straightforward at all. Take two attributes X and Y with Gaussian distributions conditional on the class variable C; suppose that a Naive Bayes classifier is learned, but one of the Gaussian distributions displays a correlation between X and Y. Depending on the value of this correlation, the asymptotic behavior for $\lambda = 1$ can be better or worse than for $\lambda \to 0$ (details can be found in (Cohen et al., 2004) and (Cozman et al., 2003)). Thus labeled data do not always produce an asymptotic classifier with better performance than unlabeled data.

Still, is there anything that can be said about the (intuitively plausible) more valuable status of labeled data? This is a key question for generative semi-supervised learning: performance degradation occurs exactly when the supervised starting point is better than the limiting unsupervised point. In the remainder of this appendix we explore this theme, focusing on maximum likelihood generative methods.

We base our argument using bounds on the classification error. For a given θ, define $EKL(\theta) = E[\log(P(C|\mathbf{X})/P(C|\mathbf{X},\theta))]$ to be the expected Kullback-Leibler divergence between the "true" posterior and the estimated posterior (Cover & Thomas, 1991). Smaller expected Kullback-Leibler divergence typically leads to smaller classification error — in the sense that expected Kullback-Leibler divergence bounds the classification error (Garg & Roth, 2001; Cover & Thomas, 1991). We now show that the expected Kullback-Leibler divergence for θ_u^* is larger than for θ_l^*, therefore the bound on the classification error with labeled data is smaller than that with unlabeled data.

Unsupervised learning asymptotically takes us to $\theta_u^* = \arg\max_\theta D_u(\theta)$, where $D_u(\theta) = E[\log P(\mathbf{X}|\theta)]$. Supervised learning instead takes us to $\theta_l^* = \arg\max_\theta D_l(\theta)$, where $D_l(\theta) = E[\log P(C, \mathbf{X}|\theta)]$. Thus $D_l(\theta) = D_t(\theta) + D_u(\theta)$, where $D_t(\theta) = E[\log P(C|\mathbf{X}, \theta)]$. Define $\theta_t^* = \arg\max_\theta D_t(\theta)$. Clearly $D_t(\theta_u^*) \leq D_t(\theta_t^*)$; simple manipulations lead to

$$EKL(\theta_u^*) \geq EKL(\theta_t^*), \tag{1}$$

In terms of expected Kullback-Leibler divergence, we have:

- Unlabeled data asymptotically yields $EKL(\theta_u^*)$ where $\theta_u^* = \arg\max_\theta D_u(\theta)$.
- Labeled data asymptotically yields $EKL(\theta_l^*)$ where $\theta_l^* = \arg\max_\theta D_t(\theta) + D_u(\theta)$.
- Direct minimization of expected Kullback-Leibler divergence yields $EKL(\theta_t^*)$ where $\theta_t^* = \arg\max_\theta D_t(\theta)$.

Note the following pattern: while unlabeled data produces an estimate by maximizing $D_u(\theta)$ (a quantity in principle unrelated to the expected Kullback-Leibler divergence), labeled data produces an estimate by maximizing a sum of $D_u(\theta)$ with the quantity of direct interest ($D_t(\theta)$).

Now return to Expression (1). If equality is attained in this expression, θ_u^* maximizes $D_t(\theta)$ and $D_u(\theta)$ simultaneously, and consequently $\theta_u^* = \theta_l^*$ where $\theta_l^* = \arg\max_\theta D_l(\theta)$. Thus equality in Expression (1) produces a situation where labeled and unlabeled data converge to the same point and have the same asymptotic value. The more interesting case where $\theta_u^* \neq \theta_l^*$ thus implies that Expression (1) is a strict inequality:

$$EKL(\theta_u^*) > EKL(\theta_t^*). \tag{2}$$

Thus θ_u^* is not a maximum of $D_t(\theta)$. We make the additional assumption that $D_t(\theta_u^*)$ is not a stationary point of $D_t(\theta)$.

At θ_u^*, the derivative of $D_l(\theta)$ is equal to the derivative of $D_t(\theta)$ (as $D_u(\theta)$ attains a maximum at θ_u^*). The maximization of $D_l(\theta)$ cannot have θ_u^* as a maximizing point, and instead will necessarily move in the direction of higher $D_t(\theta)$. Hence $D_l(\theta_u^*)$ is smaller than $D_t(\theta_l^*)$, and $EKL(\theta_u^*) > EKL(\theta_l^*)$. Consequently, the bound on the classification error is tighter with labeled data then it is with unlabeled data, suggesting that supervised learning should typically lead to better asymptotic behavior than unsupervised learning.

Obviously the analysis cannot state that labeled data are always superior to unlabeled data — this is not true, as illustrated by the Gaussian example. What the previous arguments show is that labeled data are typically more valuable than unlabeled data in asymptotically reducing a bound on classification error. Thus if the asymptotic behavior of labeled data is different from the asymptotic behavior of unlabeled data, we should expect the first to be better than the second with respect to classification. These are exactly the situations where unlabeled data may degrade the performance of generative semi-supervised learning.

Asymptotic analysis can provide insight into complex phenomena, but finite sample effects are also important. In practice one may have very little labeled data, and the estimates $\hat{\theta}$ from labeled data may be so poor that the addition of unlabeled data is a positive move. This should explain at least partially the success of generative semi-supervised learning in problems with many attributes, because in those settings the number of labeled samples is often insufficient to obtain good estimates (due to the large number of parameters involved). Text classification is an important problem where many attributes are often available, and where generative semi-supervised learning has been successful (Nigam et al., 2000).

References

Ando, R. K., & Zhang, T. (2005). *A framework for learning predictive structures from multiple tasks and unlabeled data*. IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA.

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

Belkin, M., & Niyogi, P. (2004). Semi-supervised learning on Riemannian manifolds. *Machine Learning*, *56*, 209–239.

Berk, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Annals of Mathematical Statistics*, 51–58.

Blake, C., & Merz, C. (1998). UCI repository of machine learning databases [`www.ics.uci.edu/∼mlearn/MLRepository.html`]. University of California, Irvine, Dept. of Information and Computer Sciences.

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

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

Castelli, V., & 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*, 2102–2117.

Cawley, G. C. (2000). MATLAB support vector machine toolbox (v0.50β) [`http://theoval.sys.uea.ac.uk/~gcc/svm/toolbox`]. University of East Anglia, School of Information Systems, Norwich, Norfolk, U.K. NR4 7TJ.

Cohen, I., Cozman, F., Sebe, N., Cirelo, M. C., & Huang, T. (2004). Semisupervised learning of classifiers: Theory, algorithms, and their application to human-computer interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*, 1553–1568.

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley and Sons.

Cozman, F. G., & Cohen, I. (2002). Unlabeled data can degrade classification performance of generative classifiers. *Proc. of the Fifteenth Int. Florida Artificial Intelligence Research Society Conf.* (pp. 327–331). Pensacola, Florida.

Cozman, F. G., Cohen, I., & Cirelo, M. C. (2003). Semi-supervised learning of mixture models. *Int. Conf. on Machine Learning* (pp. 99–106).

DeGroot, M. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.

Devroye, L., Gyorfi, L., & Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. New York: Springer Verlag.

Garg, A., & Roth, D. (2001). Understanding probabilistic classifiers. *European Conf. on Machine Learning (ECML)* (pp. 179–191).

Ghani, R. (2002). Combining labeled and unlabeled data for multiclass text categorization. *Int. Conf. on Machine Learning (ICML)*.

Grandvalet, Y., & Bengio, Y. (2004). Semi-supervised learning by entropy minimization. *Neural and Information Processing Systems (NIPS)*.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2003). *The Elements of Statistical Learning*. Springer.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. of the Fifth Berkeley Symposium in Mathematical Statistics and Probability*, 221–233. University of California Press.

Hyvarinen, A. (1999). Survey on independent component analysis. *Neural Computing Surveys*, *02*, 94–128.

Hyvarinen, A., & Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation*, *9(7)*, 1483–1492.

Jaakkola, T. S., Meila, M., & Jebara, T. (1999). Maximum entropy discrimination. *Neural Information Processing Systems 12*.

Joachims, T. (1999). Transducive inference for text classification using support vector machines. *Int. Conf. on Machine Learning (ICML)*.

L'Heureux, P.-J., Carreau, J., Bengio, Y., Delalleau, O., & Yue, S. Y. (2004). Locally linear embedding for dimensionality reduction in QSAR. *Journal of Computer-Aided Molecular Design*, *18*, 475–482.

McCallum, A. K. (1996). Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. [`http://www.cs.cmu.edu/~mccallum/bow`].

Miller, D. J., & 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.

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

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

Redner, R. A., & Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, *26*, 195–239.

Shahshahani, B. M., & 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*, 1087–1095.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley.

Zhang, T., & Oles, F. (2000). A probability analysis on the value of unlabeled data for classification problems. *Int. Joint Conf. on Machine Learning* (pp. 1191–1198).