A Probability Analysis on the Value of Unlabeled Data for Classification Problems

Tong Zhang TZHANG@WATSON.IBM.COM
Frank J. Oles OLES@WATSON.IBM.COM
Mathematical Sciences Department, IBM T.J. Watson Research Center, Yorktown Heights, NY 10598 USA

Abstract

Recently, there has been increasing interest in using unlabeled data for classification. In order to understand better the value of using unlabeled data under certain learning models, it is worthwhile to formulate the problem precisely, and to analyze relevant issues carefully. In this paper, we approach this problem from the statistical point of view, where we know a correct model of the underlying distribution. Fisher information matrices are used to judge the asymptotic value of unlabeled data. We apply this methodology to both "passive partially supervised learning" and "active learning", and draw conclusions. Experiments will also be provided.

1. Introduction

Many applications require classifiers built by machine learning to categorize incoming data automatically. Usually a supervised learning algorithm, requiring training data labeled by a human, is used to obtain such a classifier. Since in many applications, an enormous amount of unlabeled data is available with little cost, it is natural to ask whether, in addition to human labeled data, one can also take advantage of the unlabeled data.

There are two existing approaches to this problem. In the first approach, one trains a classifier(s) based on the labeled data as well as unlabeled data. Typically, the labels for unlabeled data are imputed by certain means based on the current state of the classifier(s). The now augmented labeled data is then used to retrain the classifier(s). Two key issues in this approach are how to impute labels of unlabeled data and how to use the augmented labeled data to retrain the classifier(s). Examples of this approach include Blum and Mitchell (1998), Castelli and Cover (1996), Joachims (1999), Nigam et al. (2000), and Shahshahani and Landgrebe (1994). In the second approach, one trains a classifier(s) based only on the labeled data. Then based on the current state of the classifier(s), one selects some of the "most informative" data so that knowing labels of the selected data is likely to greatly enhance the construction of the classifier(s). The selected data will then be labeled by a human or an oracle, and be added to the training set (to retrain the classifier(s)). This procedure can be repeated, and our goal is to label as little data as possible to achieve a certain performance. Examples of this approach include Cohn, Ghahramani, and Jordan (1997), Freund et al. (1997), McCallum and Nigam (1998), and Seung, Opper, and Sompolinsky (1992). This second approach is usually called *active learning* in the literature. In order to distinguish from it, we shall thus call the first approach *passive partially supervised learning* in this paper.

Although there have been many studies on enhancing classification performance by using unlabeled data, the existing efforts are mostly related to mixture models and ensemble methods in one form or another. In particular, there has been little analysis of the value of unlabeled data under a relatively general learning model, i.e., whether the unlabeled data can be helpful at all, and more importantly, how much it helps and what are the underlying characteristics of the model that determine the usefulness of unlabeled data. This paper addresses some aspects of this question under a probabilistic framework. Since this work is motivated from our research on text document categorization, where an enormous amount of unlabeled data is available with little cost, it is therefore natural for us to provide experiments on text-categorization problems in order to illustrate the theoretical analysis.

2. Problem Formulation

For clarity, we shall only discuss binary classification problems: that is, we would like to predict the label

$y \in \{-1, 1\}$ for a given data x. We view this problem in a probabilistic framework, where we would like to find a distribution parameter α so that the joint distribution is $p(x, y) = p(x, y|\alpha)$. The effect of unlabeled data on the efficiency of parameter estimation will be analyzed using statistical methods. As we shall see later, in this context, it is very important to distinguish the following two types of joint probability distribution models:

type 1 parametric: $p(x, y|\alpha) = p(x|\alpha)p(y|x, \alpha)$, where both $p(x|\alpha)$ and $p(y|x, \alpha)$ have known functional forms, and where $p(x|\alpha)$ has a non-trivial dependency on α.

type 2 semi-parametric: $p(x, y|\alpha) = p(x)p(y|x, \alpha)$, where the conditional probability $p(y|x, \alpha)$ still has a known functional form, but the data probability $p(x)$, decoupled from $p(y|x, \alpha)$, can have an unknown (or non-parametric) functional form independent of α.

Models of type 1 include mixture models such as mixtures of Gaussians and naive Bayesian models. The latter have been intensively applied to text categorization with reasonable results:

$$\begin{aligned} p(x, y|\alpha) &= p_y p(x|\alpha_y) \\ p(x|\alpha) &= p_{-1}p(x|\alpha_{-1}) + p_1 p(x|\alpha_1), \end{aligned}$$

where -1 and 1 are used to represent the classes. Models of type 2 include the logistic model:

$$p(x, y|\alpha) = (1 + \exp(-\alpha^T xy))^{-1} p(x), \qquad (1)$$

where the functional form of $p(x)$ is non-important. In theory, one can use the *maximum-likelihood estimate* (MLE) to determine the model parameter:

$$\hat{\alpha} = \arg\min_{\alpha} E_n \ln(1 + \exp(-\alpha^T xy), \qquad (2)$$

where E_n indicates the empirical expectation over n observed data. In practice, the MLE formulation is ill-conditioned. A standard solution is the following regularized logistic regression:

$$\hat{\alpha} = \arg\min_{\alpha} E_n \ln(1 + \exp(-\alpha^T xy) + \lambda\alpha^2. \qquad (3)$$

For text categorization, our study indicates that the regularized logistic regression achieves a performance comparable to the linear support vector machine (Dumais et al., 1998), which is generally considered as a state of the art method. This is actually not surprising since logistic regression and support vector machines have very similar loss functions.

Due to the recent popularization of SVM, it is desirable for us to analyze it in the probabilistic framework. In this paper, we use the logistic model as an approximate probability model for SVM. Our analysis and conclusions on logistic regression can then be applied to SVM. Although there are different ways to modify an SVM as a normalized probability model, and some might have a weak α dependency in $p(x)$. In our opinion, the dependency is non-essential, and it is useful to relate a SVM to a probability model of type 2 in order to understand its behavior.

3. Asymptotic Efficiency

In this paper, we judge the value of unlabeled data by evaluating its impact on the efficiency of parameter estimation. It is well-known from the standard Cramér-Rao lower-bound that for any unbiased estimator t_n of α based on n i.i.d. samples from $p(x, y|\alpha)$, the covariance of t_n satisfies $\mathrm{cov}(t_n) \geq \frac{1}{n} I(\alpha)^{-1}$, where

$$I(\alpha) = -\int p(x, y|\alpha) \frac{\partial^2}{\partial \alpha^2} \ln p(x, y|\alpha) dx dy$$

is the Fisher information matrix. Since (under quite general conditions) the maximum likelihood estimate achieves this lower bound and is unbiased asymptotically, therefore maximum likelihood estimate is the asymptotically most efficient (unbiased) estimator. Its efficiency can be measured by the Fisher information that is intrinsic to the probability model.

In the following, our discussion emphasizes the design of appropriate maximum likelihood estimates using the full information of unlabeled data. Accordingly, the value of unlabeled data can be evaluated by the gain on the corresponding Fisher information matrices. Note that this specific analysis does not catch the different behavior of the non-regularized logistic regression (2) from its regularized version equation (3). However, it is possible to generalize the analysis by either using a Bayesian approach, where we regard the regularization term as a prior, or using the traditional ill-posed system approach, where the data space of x can be infinite dimensional and the inverse of the Fisher information operator $I(\alpha)$ is considered unbounded. In this abstract, we shall only consider the standard MLE/Fisher information analysis for clarity. However, the conclusions of the extended analysis remain the same as those from the MLE analysis. The reason is that the most important difference comes from the type (1 or 2) of the probability model — a data independent prior does not change a model's type.

Even though we use the Fisher information argument

to draw conclusions, the analysis itself should only be regarded as a guide that reveals important characteristics of the underlying model assumption that have significant impact on the value of unlabeled data. This indicates that the characteristics of the model assumption revealed by the Fisher information analysis can provide valuable insights even when we only have an approximate probability model. We shall mention that the Fisher information argument has also been applied by Shahshahani and Landgrebe (1994) to study passive partially supervised learning. However, their derivation was very vague, and there was confusion about asymptotic results versus small sample results as well as confusion about the data generation mechanism. In addition, the functional form of Fisher information associated with unlabeled data was not even given by Shahshahani and Landgrebe (1994). Consequently, there exist some loopholes in their arguments (see Nigam et al., 2000).

4. Passive Partially Supervised Learning

In this section, we derive a maximum likelihood estimate that utilizes the unlabeled data, and we compute its Fisher information. The value of unlabeled data can then be quantitatively evaluated as the gain on the Fisher information. Throughout this paper, we shall assume that our model has a finite positive definite Fisher information and the appropriate MLE is both consistent and Fisher efficient, which is valid under quite mild assumptions. [1]

In order to obtain an efficient MLE, we shall consider the following model of data generation. There is an unknown ratio $\gamma \in [0,1]$ which is drawn from an unknown distribution $P(\gamma)$. We draw n independent samples x: with probability γ, we give it a known label $y \in \{-1,1\}$; with probability $1-\gamma$, the label is unknown. In the case of unknown label, we identify the data with $y=0$. Now, the joint data distribution is a mixture of

$$p(x, y=\pm 1|\alpha) = \int p(x,y|\alpha)\gamma dP(\gamma) = p(x,y|\alpha)\bar{\gamma}$$

and

$$p(x, y=0|\alpha) = \int p(x|\alpha)(1-\gamma)dP(\gamma) = p(x|\alpha)(1-\bar{\gamma}),$$

where $\bar{\gamma} = \int \gamma dP(\gamma)$ is the expectation of γ.

For a probability model of type 1, we now assume that an oracle knows $\bar{\gamma}$. With this knowledge, the asymptotically most efficient estimator is MLE which becomes

$$\hat{\alpha}_{\bar{\gamma}} = \arg\sup_{\alpha} \sum_i \ln[p(x_i, y_i|\alpha)\bar{\gamma}] + \sum_j \ln[p(x_j|\alpha)(1-\bar{\gamma})],$$

where the index i goes over labeled data and the index j goes over unlabeled data. This asymptotically most efficient estimator of α under the assumption of knowing the extra knowledge of $\bar{\gamma}$ is exactly the same estimate as

$$\hat{\alpha} = \arg\sup_{\alpha} \sum_i \ln p(x_i, y_i|\alpha) + \sum_j \ln p(x_j|\alpha)$$

of α without knowing $\bar{\gamma}$. The Fisher information of this estimator (which depends on $\bar{\gamma}$) is given by

$$\begin{aligned} & I_{labeled+unlabeled}(\alpha) \\ = & -\bar{\gamma}\int p(x,y|\alpha)\frac{\partial^2}{\partial\alpha^2}\ln p(x,y|\alpha)dxdy \\ & -(1-\bar{\gamma})\int p(x|\alpha)\frac{\partial^2}{\partial\alpha^2}\ln p(x|\alpha)dx \\ = & I_{labeled}(\alpha) + I_{unlabeled}(\alpha). \end{aligned}$$

Since for models of type 1, when $\bar{\gamma} < 1$, the Fisher information $I_{unlabeled}(\alpha)$ is non-zero, therefore including unlabeled data always helps.

For a semi-parametric probability model of type 2, we consider the maximum likelihood estimate corresponding to an oracle that knows the precise distribution $p(x)$ as well as $\bar{\gamma}$. By arguments similar to those outlined above, this optimal MLE is the same as the following estimator without any knowledge of either $p(x)$ or $\bar{\gamma}$:

$$\hat{\alpha} = \arg\sup_{\alpha} \sum_i \ln p(y_i|\alpha, x_i),$$

where index i goes over labeled data. The corresponding Fisher information is

$$\begin{aligned} & I_{labeled+unlabeled}(\alpha) \\ = & -\bar{\gamma}\int p(x,y|\alpha)\frac{\partial^2}{\partial\alpha^2}\ln p(y|\alpha,x)dxdy \\ = & I_{labeled}(\alpha). \end{aligned}$$

This indicates that for models of type 2, unlabeled data does not help (at least asymptotically). This conclusion is not surprising since for a model of type 2, the data distribution $p(x)$ does not carry any parameter information. Therefore including data points without labels clearly won't have any effect on parameter estimation.

[1] A simple well-known example for the inconsistency of MLE is the mixture model density estimation allowing the variance of a mixture component to approach 0. In this case, an MLE can over-fit any particular data point leading to an infinite likelihood. Such pessimistic models will be excluded in this paper.

Due to the relationship between the logistic regression, which is a probability model of type 2, and the support vector machine, our analysis implies that transductive SVM (Joachims, 1999; Wu et al., 1999) in its current form is unlikely to be helpful in general. This agrees with the PAC analysis by Wu et al. (1999), which also suggests that transductive SVM is not very helpful. However, this opinion contradicts some other studies, most noticeably Joachims (1999). Therefore we would like to investigate this issue further.

In order for unlabeled data to have an impact on the parameter estimation, the data distribution $p(x)$ should be parameter dependent. In the case of logistic regression and SVM, a strong parameter dependency of $p(x)$ is not necessary for these methods to work well in the supervised setting. The success of these methods only indicates that there exists a reasonably large margin between in-class and out-of-class members. However, in the passive partially supervised setting, the basic data distribution assumption of a transductive SVM is that $p(x)$ should have an artificial margin that is as large as possible, so that labels can be imputed according to this artificial margin. There is insufficient evidence so far (both in theory and in practice) to indicate that this artificial margin indeed has much to do with the true class separation, especially for problems containing multiple clusters (and thus multiple possible large margin separations).

We implemented a version of transductive support vector machine and applied it to text categorization investigated by Joachims (1999). We use the Mod-Apte split of the Reuters-21578 data set availabel from *http://www.research.att.com/~lewis/reuters21578.html*. In our experiments, we use word stemming without any stop-word removal or feature selection. Although in some specific setups, there might be some improvements (especially if the parameters are tuned in favor of transduction), we have found no statistically significant evidence that transduction is helpful in the general situation.

To understand our experiments better, we report the result from one specific run over the category "earn" in Reuters. We randomly selected 20 data points from the 9603 in the training data of the Mod-Apte split as labeled data. The selection contained 5 in-class members and 15 out-of-class members. We used the remaining training data (9583 data points) as unlabeled data to train a transductive SVM. Results are depicted in Figure 1. The top row in Figure 1 contains the histograms of the projection (by inner products) of the 9603 training data to the computed linear classifier weight from the supervised SVM (using labeled data only), where the projections of the in-class data, out-of-class data, and the combination of the two are plotted separately. The middle row in Figure 1 contains the corresponding histograms with the weight computed from a transductive SVM (using both labeled and unlabeled data). The bottom row contains the scatter plots of the unlabeled in-class data, unlabeled out-of-class data, as well as the labeled data, where the x-axis is the projection to the supervised SVM weight and the y-axis is the projection to the transductive SVM weight. In the bottom right scatter plot, each labeled in-class data is marked by a triangle and each labeled out-of-class data is marked by a circle. Note that the labeled data were perfectly separated with both the supervised SVM and the transductive SVM.

Taking the perfect separation of labeled data into consideration, without any prior knowledge of the labels for unlabeled data, the histogram from the transductive SVM looks much more appealing since there is a significant margin that separates two Gaussian like components for the unlabeled data. Unfortunately, this large margin achieved by the transductive SVM was accomplished by pushing many (unlabeled) in-class data to the wrong direction. In fact, the generalization performance of the transductive SVM evaluated on the unlabeled data is worse than that of the supervised SVM despite the seemingly more appealing unlabeled margin distribution. It is now clear that, in practice, the standard transductive argument may mislead the classifier into maximizing the "wrong margin". To our knowledge, this issue has not yet been addressed in any of the current proposed forms of using an SVM like classifier for passive partially supervised learning. This suggests that the success reported in the literature might be due to their specific experimental setups rather than the general advantage of a transductive SVM versus a supervised SVM. In order to take advantage of unlabeled data for a discriminative model, it is necessary to impose a generally suitable parameter dependent data model $p(x)$, which is still not available yet. Our experiments suggest that margin maximization by itself seems to be unreliable.

5. Active Learning

While probability models of type 1 are suitable for passive partially supervised learning, probability models of type 2 are suitable for active learning. This is because the consistency of a parameter estimation procedure for the latter does not depend on the data distribution $p(x)$, while the efficiency can vary with different choices of $p(x)$. On the other hand, it is only possible to apply active learning to probability models of type 1

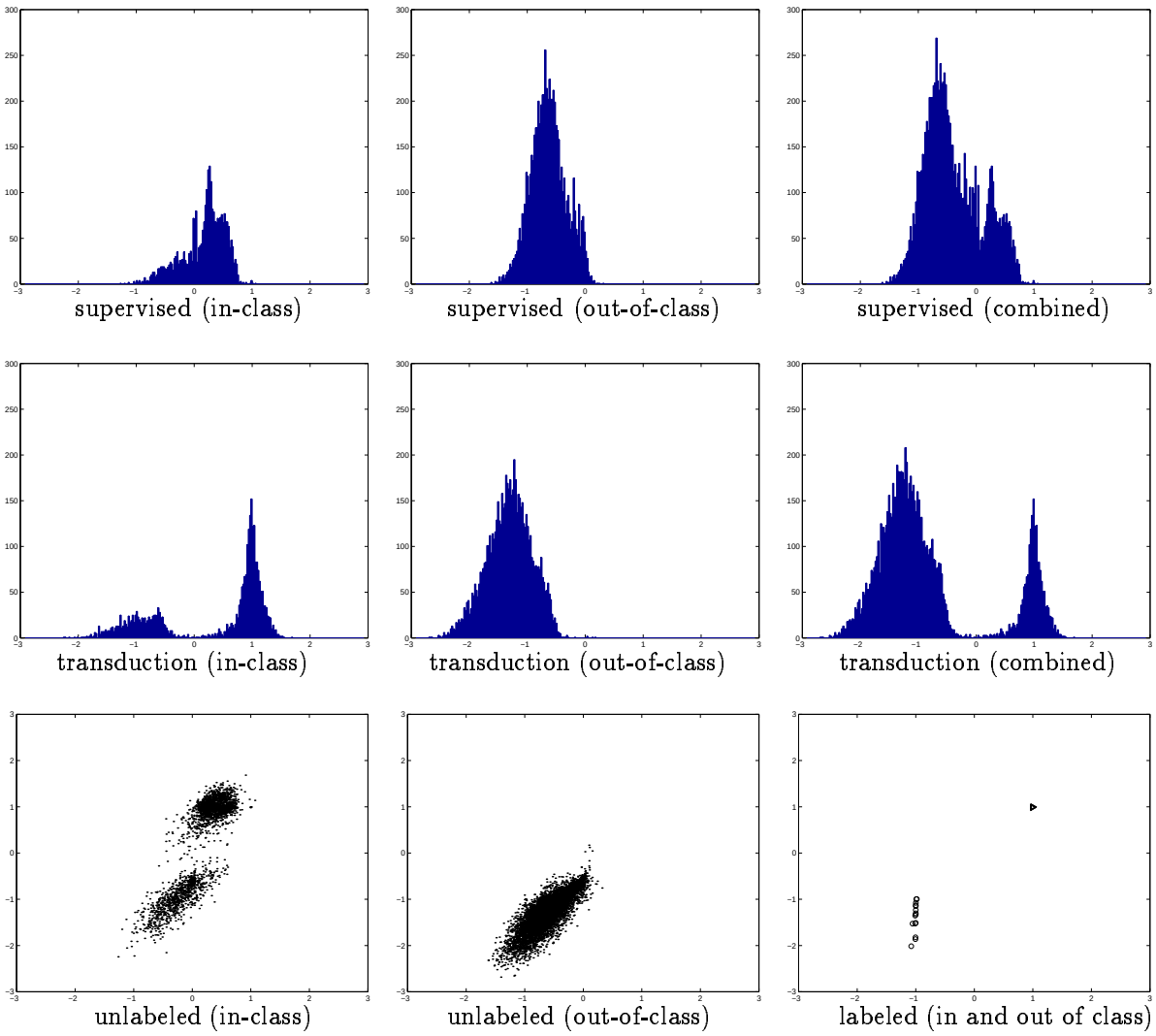

Figure 1. Supervised SVM (20 labeled data) vs. transductive SVM (20 labeled + 9583 unlabeled data) on the Reuters "earn" category.

indirectly, since a change of $p(x)$ may affect or violate the model assumption. However, by using a sufficient amount of unlabeled data, we can eliminate the part of parameter α that is $p(x)$ dependent. Active learning can then be applied to determine the part of the parameter that is invariant to a change of $p(x)$.

To analyze active learning for probability models of type 2, we shall consider a resample $q(x)$ of the unlabeled data so that the asymptotic efficiency of estimating α measured by the Fisher information

$$I_q(\alpha) = -\int q(x)dx \int p(y|\alpha,x)\frac{\partial^2}{\partial\alpha^2}\ln p(y|\alpha,x)dy$$

is "maximized". One can use the mean squared error of the estimated parameter to measure the goodness of q, but this is often not fully correlated with the classification error. Although the expected classification error itself can be used, it leads to a more complicated form than the expected log-likelihood which is asymptotically given by (the proof will be skipped):

$$E_{X_1^n}\int p(x)dx \int \ln\frac{p(y|\hat{\alpha}_n,x)}{p(y|\alpha,x)}p(y|\alpha,x)dy$$
$$= -\frac{1}{2n}\mathrm{tr}(I_q(\alpha)^{-1}I_p(\alpha)),$$

where $E_{X_1^n}$ is the expectation over n samples from $q(x)$; $\hat{\alpha}_n$ is the MLE; α denotes the true parameter; I_p and I_q denotes the Fisher information with respect to the original data distribution $p(x)$ and the resampled

data distribution $q(x)$ respectively. The Cramér-Rao lower-bound implies that the MLE based on the resampled distribution q that minimizes $\mathrm{tr}(I_q(\alpha)^{-1} I_p(\alpha))$ is the asymptotically most efficient estimator of α (as far as its expected log-likelihood is concerned) among all estimators based on a resampling of x.

To apply this result to active learning, we assume that we have a good estimate $\hat{\alpha}$ of α and then replace α by $\hat{\alpha}$ to estimate the optimal resampled distribution:

$$\hat{q} = \arg\inf_q \mathrm{tr}(I_q(\hat{\alpha})^{-1} I_p(\hat{\alpha})). \quad (4)$$

More samples can then be drawn from $\hat{q}(x)$ and we re-estimate $\hat{\alpha}$ as well as the optimal sample distribution $\hat{q}$. This procedure can be repeated.

In related work, Cohn et al. (1997) considered active learning with squared loss in a regression setting. Their statistical analysis based on the bias-variance trade-off is very related to our analysis based on the Fisher information. For example, they assume that the sample selection mechanism won't affect the bias which corresponds to our assumption that the probability model is of type 2. Their criterion is to minimize the variance which corresponds to the maximization of Fisher information in our analysis. Although given an exact probability model, our argument based on the Fisher information is already general, it is possible to further extend this argument to a non maximum likelihood estimate (such as a support vector machine) with a probability confidence measure.

As an example for equation (4), we consider the logistic regression, where the Fisher information is given by

$$I_q(\alpha) = \int \frac{1}{(1+e^{\alpha^T x})(1+e^{-\alpha^T x})} x x^T q(x) dx.$$

If I_q is estimated from the empirical data when the number of data points is less than the dimension, then I_q is singular. In this case, a regularization term has to be added. Another practical issue is that the optimization of (4) is usually very difficult. In this paper, we propose to identify the key factors in the optimal sampling strategy based on insights provided by the Fisher information analysis, so that equation (4) is heuristically optimized. This should work well in practice since a precise model is usually not available and hence the exact minimization of equation (4) is non-essential.

For logistic regression, to maximize the Fisher information $I_q(\alpha)$, we shall favor an unlabeled data point x so that its contribution to the Fisher information

$$\frac{1}{(1+e^{\alpha^T x})(1+e^{-\alpha^T x})} x^T x \quad (5)$$

is significant. This indicates that we prefer a data x such that its projection $\alpha^T x$ is small (margin is small) and its size is large ($x^T x$ is large). To prefer a data point that has a small margin is quite intuitive based on previous studies of committee-based algorithms such as Freund et al. (1995) and Seung et al. (1992): the label of the most uncertain data is likely to reveal most important information. To prefer a large x is less intuitive at the first glance. However, this criterion is also natural since in a logistic model, a small x (the extreme case is $x = 0$) is inherently uncertain so that its label does not reveal any useful deterministic information (for all α, the label of $x = 0$ is completely random: $P(y = \pm 1) = 0.5$). This important consideration is not an issue in the query by committee formulation (Seung et al., 1992), since they assume that perfect classification is always achievable. In general, the following two principles are implications from equation (4):

- Choose an unlabeled data point of low confidence with the estimated parameter such that it can have a potentially significant increase in confidence with the true (or re-estimated) parameter.
- Choose an unlabeled data point that shall not be redundant with other choices (or data already chosen).

As another good example to show that low confidence of a data point itself is an insufficient indicator, we consider the mixture of two one-dimensional unit-variance Gaussians with unknown centers at ± 1. Since this model is of type 1, we can use the passive partially supervised learning to obtain the centers at ± 1 (Castelli & Cover, 1996). The remaining problem, which can be achieved by active learning, is to determine which label corresponds to which center. With a flat prior knowledge, any data is completely non-confident since its label is ± 1 with probability 0.5. However, in an active learning setting, we would like to label a data point in the extreme tail of the distribution that has the greatest potential of enhancing its confidence non-randomly.

Returning to the logistic regression formulation, we study the application of active learning on text categorization problem. Again, the Reuters data set is used for illustration. We observed in our experiments that the size of x is actually less relevant than its margin $\alpha^T x$ as a criterion for good sample: using (5) rather than simply favoring data with smaller margins gives a slightly poorer performance (although both methods are significantly better than random sampling). We conjecture the following two explanations. One is that

the effect of x has to be discounted by $I_p(\hat{\alpha})$ in equation (4), which we do not consider in the heuristics.[2] Another reason is that the logistic model assumption is only approximate for text categorization problems, and hence using margin is more robust than using (5) which requires the exact validity of the logistic model.

For active learning, we start with 100 randomly chosen labeled samples. We then use the margin criterion to pick more samples to label: each time, the sample size is increased by 50% (up to the predetermined sample size to be labeled). The parameter is then re-estimated, and the procedure repeated until the predetermined label size is achieved. We compare this scheme with randomly chosen samples. In text categorization, the performance is usually measured by precision and recall rather than classification error.

$$\text{precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \times 100,$$

$$\text{recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \times 100.$$

For a linear classifier that has a threshold parameter to trade-off the precision and the recall, it is typical to report the break-even point where precision equals recall. Since a document in the Reuters dataset can be multiply categorized, it is common to study the dataset as separate binary classification problems, where each problem corresponds to a category. The overall performance can be measured by the micro-averaged precision, recall and the break-even point computed from the overall confusion matrix defined as the sum of individual confusion matrices corresponding to the categories.

We use the top ten categories (the remaining categories are typically very small) for this study. Note that for active learning, the sample selection mechanism is based on each individual binary classification problem. This is sufficient for the purpose as a demonstration of principle for our analysis. Unfortunately, we do not have any space to report the performance of a practically more sensible strategy in which the same selected samples are used for all categories.

Figure 2 compares the performance of active learning vs. random sampling measured by micro-averaged break-even points as a function of labeled training samples, evaluated on the standard Mod-Apte test set which is consisted of 3299 documents. Each data point in the plot corresponds to ten random runs. The center is the mean, and the size of the error-bar is the standard deviation. The break-even point achieved by logistic regression with all 9603 training data is 91.9 which is comparable with an enhanced SVM (Dumais et al., 1998). For active learning, this performance is already achieved with about 1000 samples. As a comparison, with even 5000 random samples, the performance of 91.9 is not yet achieved. Also note that active learning tends to give a smaller variance for two reasons: it tends to select some fixed informative samples; and the performance of active learning in our model is achieved through variance reduction. A more detailed comparison also shows that active learning out-performs random sampling for each of the ten categories we have studied.

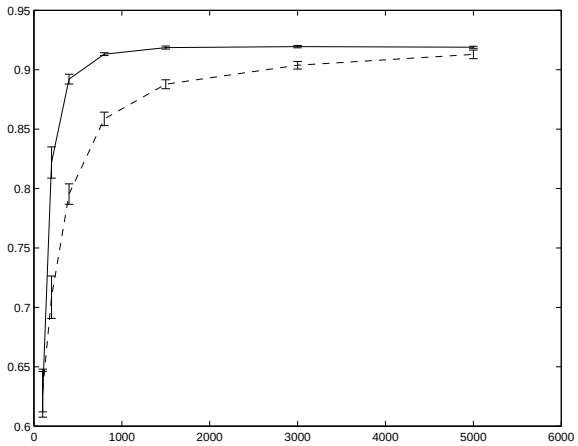

Figure 2. Break-even points of logistic regression as a function of labeled sample size: active learning = "solid"; random sampling = "dashed".

6. Discussion

Although we have emphasized classification problems, the analysis is also suitable for other learning problems where we want to predict a certain variable y based on an observed variable x. In all such cases, it is important to distinguish probability models of type 2 from probability models of type 1. Specifically, probability models of type 1 are suitable for passive partially supervised learning while probability models of type 2 are suitable for active learning. Intuitively, a probability model of type 1 tends to be a generative model in that each class parameter is defined by class members alone. A probability model of type 2 tends to be a discriminative model in that the model parameter is not

[2] For example, if a component x_j of x is irrelevant, so that $\alpha_j = 0$, then x_j should not be counted in the dot product $x^T x$. Note that this is automatically discounted in $\alpha^T x$.

for the purpose of generating the class members, but rather of discriminating in-class members from out-of-class members.

A specific but important conclusion from our analysis is that support vector machines in their current forms are not suitable for passive partially supervised learning. Although this seems to contradict some previous claims, we believe that the earlier reported success might be due to specific experimental set-ups. In particular, the issue of "maximizing the wrong margin" was not addressed in any previous approach of using unlabeled data for passive partially supervised learning with an SVM-like classifier. We believe that it is important to carefully analyze the previous studies in order to understand how to avoid this phenomenon of "maximizing the wrong margin". If we can indeed identify some key factors that helped those experiments to alleviate the problem we have encountered, then the current standard approach can be reformulated in a more appropriate form so that real progress can be made.

It is also important to note that from our analysis, support vector machines are suitable for active learning in its current form. The very reason that active learning works for SVM also supports the claim that passive partially supervised learning is not suitable for SVM. This is because an unlabeled data point with a small margin is likely to cause a large change in parameter estimation once its label is known. The label itself is intrinsically non-deterministic, therefore an attempt to push such a data point to a deterministic state (e.g., by margin maximization) will fail unless a better probability model containing information not captured by the Fisher information in the current SVM model is used. This argument again demonstrates why it is important to refine the standard SVM so as to use it for passive partially supervised learning.

References

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. *Proceedings of the Eleventh Annual Conference on Computational Learning Theory* (pp. 92–100). New York: ACM.

Castelli, V., & Cover, T. (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.

Cohn, D., Ghahramani, Z., & Jordan, M. (1997). Active learning with mixture models. In Murray-Smith, R., & Johanson, T. (Eds.), *Multiple model approaches to modelling and control*, 167–183. Taylor & Francis.

Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. *Proceedings of the Seventh ACM International Conference on Information and Knowledge Management* (pp. 148–155). New York: ACM.

Freund, Y., Seung, H., Shamir, E., & Tishby, N. (1997). Selective sampling using the query by committee algorithm. *Machine Learning, 28*, 133–168.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. *Proceedings of the Sixteenth International Conference on Machine Learning* (pp. 200–209). San Francisco: Morgan Kaufmann.

McCallum, A., & Nigam, K. (1998). Employing em in pool-based active learning for text classification. *Proceedings of the Fifteenth International Conference of Machine Learning* (pp. 350–358). San Francisco: Morgan Kaufmann.

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. *Machine Learning, 39*, 1–32.

Seung, H., Opper, M., & Sompolinsky, H. (1992). Query by committee. *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory* (pp. 287–294). New York: ACM.

Shahshahani, B., & Landgrebe, D. (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.

Wu, D., Bennett, K. P., Cristianini, N., & Shawe-Taylor, J. (1999). Large margin decision trees for induction and transduction. *Proceedings of the Sixteenth International Conference on Machine Learning.* (pp. 474–483). San Francisco: Morgan Kaufmann.