The Value of Unlabeled Data for Classification Problems

Tong Zhang
Mathematical Sciences Department,
IBM T.J. Watson Research Center,
Yorktown Heights, NY 10598 USA
tzhang@watson.ibm.com

Abstract

Recently, there has been increasing interest in using unlabeled data for classification. However, whether these unlabeled data are truly useful is still under debate. In order to have a better understanding of relevant issues, it is worthwhile to precisely formulate the problem and carefully analyze the value of unlabeled data under certain learning models. In this paper, we approach this problem from the statistical point of view, where we assume that a correct model of the underlying distribution is given. We demonstrate that Fisher information matrices can be 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 from this analysis. Experiments will be provided to support our claims.

1 Introduction

In today's world, an enormous amount of information is available in electronic form. In order to process these data, it is very useful to organize them so that similar data are grouped together. It is also very desirable that the data can be organized automatically by a computer program. This leads to a classification problem. Typically, a human has to set up the categories and assign labels to each data point. A supervised machine learning algorithm will then be employed to construct an underlying classification rule from the labeled data so that future unlabeled data can be automatically categorized. In order to obtain a desirable machine-constructed categorizer under this scenario, the required human labeling effort can be extremely tedious and time consuming. It is thus very important to reduce this human labeling effort as much as we can.

Since in many applications, enormous amounts of unlabeled data are available with little cost, it is therefore natural to ask the question that in addition to human labeled data, whether one can also take advantage of the unlabeled data in order to improve the effectiveness of a machine-learned categorizer.

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 label of an unlabeled data point is 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 are [1, 2, 7, 10, 13, 12]. The second approach does not impute labels for the unlabeled data in the training phase. Instead, one first 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 are [3, 5, 8, 9, 11]. 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 previous 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 on the value of unlabeled data under a relatively general learning model, *i.e.* whether the unlabeled data can be truly helpful at all (under a certain learning model), and more importantly, how much it helps and what is the underlying characteristics of the model that determines the usefulness of unlabeled data. This paper addresses some aspects of this question under a probabilistic framework. Although we do not intend to provide a direct solution under other learning models, our analysis provides valuable insights into those methods so that the usefulness of unlabeled data can be characterized. 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 model: $p(x, y|\alpha) = p(x|\alpha)p(y|x, \alpha)$, where both $p(x|\alpha)$ and $p(y|x, \alpha)$ have known functional forms. $p(x|\alpha)$ has a non-trivial dependency on α.

type 2 semi-parametric model: $p(x, y|\alpha) = p(x)p(y|x, \alpha)$, 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 where the latter have been intensively applied to text categorization with reasonable results:

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

Models of type 2 include the logistic model:

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

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:

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

where E_n indicates the empirical expectation over n observed data. In practice, the MLE formulation is ill-conditioned. It is therefore necessary to employ the following regularized logistic regression with appropriate chosen λ:

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

Recently, the regularized logistic regression has been applied to text categorization problems [14] with a performance comparable to the linear support vector machine [6, 4] which is generally considered as a state of the art method for text-categorization. This is actually not surprising since logistic regression and support vector machines have very similar loss functions — hence they should have comparable performances.

Due to the recent popularization of SVM, it is very 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. Note that there are different ways to modify an SVM as a normalized probability model. Some might have a very weak α dependency in $p(x)$ which is actually non-essential. In our opinion, it is useful to relate an 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

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

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 will emphasize 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 capture the different behavior of the non-regularized logistic regression (2) from its regularized version (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. For text-categorization problems, since the data dimension can be much more than the number of the data points, it is very reasonable to regard it as infinite dimensional so that the ill-posed system point of view becomes appropriate.

In the following, we shall only consider the standard MLE/Fisher information analysis for clarity. The extended analysis for regularized formulation will be more carefully discussed in the full paper. However, the conclusions of the extended analysis will be the same of those from the MLE analysis. One reason is that we shall see shortly that the most important difference comes from the type (1 or 2) of the probability model. Note that 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 in [12] 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 in [12]. Consequently, there exist some loopholes in their arguments (see [10]).

4 Passive partially supervised learning

In this section, we shall derive a maximum-likelihood estimate that utilizes the unlabeled data and 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].

[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.

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. Note that if $\bar{\gamma} = 0$, then α may not be fully determined even if I is positive definite at the optimal parameter. The reason is that the standard (and obvious) regularity condition for the consistency of MLE (in fact, for any estimator) that “the probability distribution by any two different values of α are distinct” may be violated for certain models (see related discussions in [2, 10]).

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 similar arguments 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

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

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.

Due to the relationship between the logistic regression which is a probability model of type 2, and the support vector machine, an important consequence from our analysis implies that transductive SVM in its current form is unlikely to be very helpful in general. This statement contradicts some earlier studies (although there are also supports for similar conclusions), most noticeably [7]. Therefore we would like to investigate this issue further.

As mentioned before, 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 $p(x)$ (without any knowledge of the data labels) should have an artificial margin that is as large as possible, so that labels can be imputed according to this artificially determined 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 separation between classes, especially for problems containing multiple clusters (and thus multiple possible large margin separations) typically observed in practice.

In order to support our analysis, we have implemented a version of transductive support vector machine and applied it to the text categorization problem investigated in [7]. We use the Mod-Apte split of the Reuters-21578 dataset 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 improvement (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 what really happened in our experiments, we report the result from one run over the category "earn" in Reuters. We select (randomly) 20 data from the 9603 training data in the Mod-Apte split to label. In the case shown in Figure 1, we have 5 in-class members and 15 out-of-class members. We use the rest 9583 training data points as unlabeled data to train a transductive version of SVM. The top line 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 by 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 line in Figure 1 contains the corresponding histograms with the weight computed from a transductive SVM by using both labeled and unlabeled data. The bottom line contains the scatter plots of the unlabeled in-class data, unlabeled out-of-class data, as well as the labeled data, where x-axis is the projection to the weight from the supervised SVM and the y-axis is the projection to the weight from the transductive SVM. 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. It can be seen that the labeled data has been 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, after we look at the true labels, it becomes clear that the large margin achieved by the transductive SVM is 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 of the seemingly more appealing unlabeled margin distribution.

It is 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 (related to probability model of type 2 which is discriminative in nature) for passive partially supervised learning. This suggests that the success reported in the literature is likely due to their specific experimental setups rather than the general advantage of a transductive SVM versus a fully 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 (unfortunate, margin maximization itself does not seem to be a very reliable data model for this purpose).

5 Active learning

Interestingly, 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 $p(x)$. It is therefore possible for us to change the data distribution to achieve a better efficiency on the parameter estimation.

On the other hand, it is only possible to apply active learning to probability models of type 1 indirectly, since a change of $p(x)$ may affect the model parameter or even violate the model assumption. However, by using a sufficient number 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 distribution $p(x)$.

In order to apply the Fisher information criterion 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". A natural question is the criterion to determine which I_q is better. One can use the mean squared error of the estimated parameter, which is often not fully correlated with the classification error. Although the expected classification error itself can be used, it often leads to a more complicated form than the expected log-likelihood which is asymptotically given by (the proof will be skipped in this abstract):

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

where E_n is the expectation over n independently resampled data from $q(x)$; $\hat{\alpha}_n$ is the maximum likelihood estimate from the resampled data; 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 maximum likelihood estimate based on the resampled distribution q that minimizes $\mathrm{tr}(I_q(\alpha)^{-1}I_p(\alpha))$ is the asymptotically most efficient parameter estimate of α (as far as its expected log-likelihood is concerned) among all estimators based on some resampling of the data distribution.

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:

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

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 a related work [3], the authors considered active learning with the squared loss in a regression setting. Interestingly they tried to apply the framework to mixture models (note that for classification problems, passive partially supervised learning should already be very suitable for such models). 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, we can further extend this argument to a non maximum likelihood estimate (such as a support vector machine) with a probability confident measure. The main reason to use a non maximum likelihood estimate is that the distribution model is usually not exact in practice, therefore the asymptotic optimality of MLE is not important. The general formulation for active learning with non maximum likelihood estimates will be given in the full paper.

As an example for (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 with the number of data points 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 problem in (4) is usually difficult to solve. Monte Carlo simulation was employed in [3]. 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 (4) is heuristically optimized. This should work well in practice since a precise model is usually not available and hence the exact minimization (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

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

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 that has a small margin is quite intuitive based on previous studies of committee based algorithms such as [5, 11]: the label of the most uncertain data is likely to reveal most important information. To prefer large x is less intuitive at the first glance. However, this criterion is also quite natural since in a logistic model, if x is small (the extreme case is $x = 0$), it is inherent uncertain so that its label does not reveal any useful deterministic information (*e.g.* 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 [11] since they assume that perfect classification is always achievable. In general, the following two principles are implications from (4):

- Choose an unlabeled data 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 that shall not be redundant with other choices (or data already chosen).

As another good example to show that low confidence of a data 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 [2]. Note that the problem has a symmetry, therefore the remaining problem is to determine which label corresponds to which center. Since the $p(x)$ dependency has been removed in the passive partially supervised learning stage, we can use active learning to determine the label correspondence for this remaining problem. 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 corresponding to the extreme tail of the joint distribution since this gives the greatest potential of enhancing its confidence non-randomly.

Returning to the logistic regression formulation: in the following, we study the performance of active learning on text categorization problem. Again, we use the Reuters dataset for illustration. It is interesting to observe from our experiments that the size of x is less relevant than its margin $\alpha^T x$ as a criterion of good sample. That is, using (5) rather than simply favoring data with smaller margins seems to give 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

(4), which we don't 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:

$$\begin{aligned} \text{precision} &= \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \times 100 \\ \text{recall} &= \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \times 100 \end{aligned}$$

Since a linear classifier contains a threshold parameter that can be adjusted to trade-off the precision and the recall, it is typical to report the break-even point, where precision equals recall. Since each document in the Reuters dataset can be multiply categorized, it is common to study the dataset as separate binary classification problems, each corresponding 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. Although this is sufficient for our purpose as a demonstration of principle for our analysis, it is not suitable for practical purposes where the sample selection mechanism should be the same for all categories. Our analysis and sample selection method can be modified to deal with such situation and still achieves a significant performance enhancement. These results will be reported in the full paper.

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 testset 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 version of support vector machine [4] (also see [14] for more comparisons). 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 due to the following two reasons: 1. it tends to select some fixed informative samples; 2. the performance of active learning in our model is achieved through variance reduction. We also list the detailed comparisons of active learning vs. random sampling for the top ten categories at 1000 labeled sample size in 1. It is clear that active learning performs consistently better than random sampling for all categories.

[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 $\alpha^T x$.

6 Discussions

In this paper, we have investigated the possibility of using unlabeled data for supervised learning under the probabilistic framework. We apply the Fisher information criterion to analyze the asymptotic value of unlabeled data when our probability model assumption is exact. However, this analysis also provides valuable insights into situations when our model assumption is not exact. This point has been illustrated in the paper through our analysis of the support vector machine.

Although we have emphasized classification problems, the analysis is also suitable for other learning problems where we want to predict certain variable y based on observed variable x (such as regression). 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 (like mixture models) 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 for the purpose of generating the class members, but rather of discriminating in-class members from out-of-class members through maximizing the log-likelihood of the conditional probability.

A specific but important conclusion from our analysis is that support vector machines in its current form are not particularly suitable for passive partially supervised learning. Our experiments confirm with this analysis. 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" observed in our experimental study was not addressed at all in any previous approach of using unlabeled data for passive partially supervised learning with a support vector machine 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 very suitable for active learning in its current form. This has also been confirmed in our experiments. Not surprisingly, 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 very likely to cause a large change in parameter estimation once its label is known. The label itself is intrinsically highly non-deterministic, therefore any attempt to push such an unlabeled 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 standard SVM model is used. This argument again demonstrates why it is important to refine the current SVM model so as to use it for passive partially supervised learning.

References

[1] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *proceedings ofthe eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

[2] V. Castelli and T.M. Cover. 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, 1996.

[3] D.A. Cohn, Z. Ghahramani, and M.I. Jordan. Active learning with mixture models. In Roderick Murray-Smith, Tor Arne Johanson, and T.A. Johansen, editors, *Multiple model approaches to modelling and control*, pages 167–183. Taylor & Francis, 1997.

[4] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the 1998 ACM 7th international conference on Information and knowledge management*, pages 148–155, 1998.

[5] Y. Freund, H.S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

[6] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In *European Conference on Machine Learing, ECML-98*, pages 137–142, 1998.

[7] Thorsten Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of International Conference on Machine Learning*, 1999.

[8] J. Krogh, A. Vedelsby. Neural network ensembles, cross validation, and active learning. In *Proceedings of NIPS'94*, pages 231–238, 1995.

[9] Andrew McCallum and Kamal Nigam. Employing em in pool-based active learning for text classification. In *Proceedings of the Fifteenth International Conferencef Machine Learning (ICML 98)*, pages 350–358, 1998.

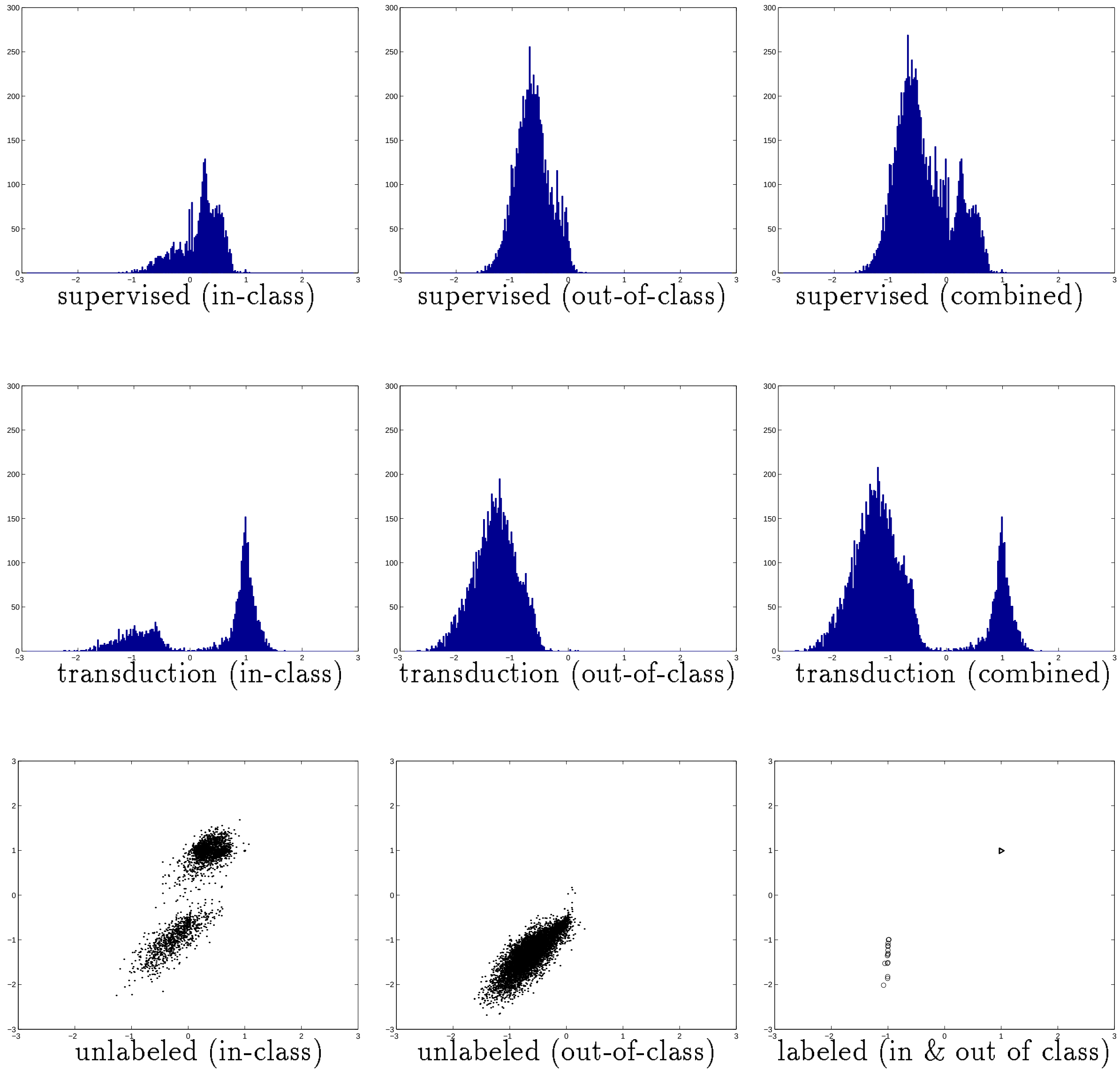
[10] Kamal Nigam, Andrew K. Mccallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39((2/3)):1–32, 2000.

[11] H.S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the fifth annual ACM workshop on Computational learning theory*, pages 287–294, 1992.

[12] B.M. Shahshahani and D.A. Landgrebe. 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, 1994.

[13] G. Towell. Using unlabeled data for supervised learning. In *Proceedings of 1995 Conference on Advances in Neural Information Processing Systems*, pages 647–653, 1996.

[14] Tong Zhang and Frank Oles. Text categorization based on logistic regression and a modification of the least square fit method. Manuscript, 2000.



supervised (in-class)

supervised (out-of-class)

supervised (combined)

transduction (in-class)

transduction (out-of-class)

transduction (combined)

unlabeled (in-class)

unlabeled (out-of-class)

labeled (in & out of class)

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

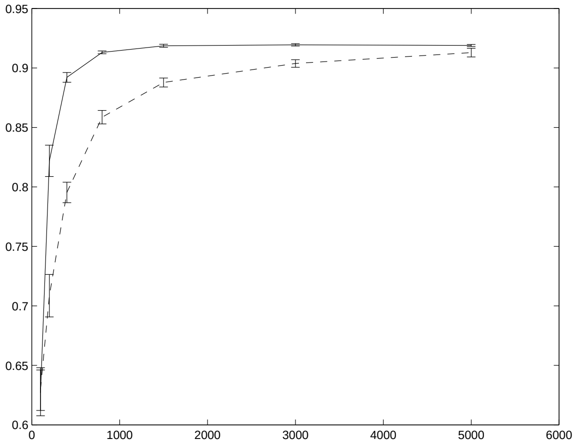


Figure 2: Break-even points of logistic regression as a function of labeled sample size: active learning = 'solid'; random sampling = 'dashed'.

category	random sampling	active learning
earn	97.7 ± 0.3	98.6 ± 0.1
acq	93.1 ± 0.7	94.9 ± 0.3
money-fx	71.8 ± 2.2	76.8 ± 1.1
grain	79.7 ± 2.4	89.7 ± 0.3
crude	83.4 ± 1.2	88.0 ± 0.5
trade	74.2 ± 3.0	76.9 ± 0.6
interest	71.3 ± 4.3	73.3 ± 1.4
ship	76.0 ± 3.2	84.8 ± 0.6
wheat	75.3 ± 3.5	82.8 ± 0.6
corn	63.9 ± 4.6	86.8 ± 0.9
micro-average	88.8 ± 0.4	91.9 ± 0.1

Table 1: Active learning vs. random sampling for the top ten Reuters categories (labeled sample size = 1000).