A New Analysis of the Value of Unlabeled Data in Semi-Supervised Learning for Image Retrieval

Qi Tian, Jie Yu, Qing Xue, Nicu Sebe*
Department of Computer Science, University of Texas at San Antonio, TX 78249
{qitian, jyu, qxue}@cs.utsa.edu
**Faculty of Science, University of Amsterdam, The Netherlands*
nicu@science.uva.nl

Abstract

There has been an increasing interest in using unlabeled data in semi-supervised learning for various classification problems. Previous work shows that unlabeled data can improve or degrade the classification performance depending on whether the model assumption matches the ground-truth data distribution, and also on the complexity of the classifier compared with the size of the labeled training set. In this paper, we provide a new analysis on the value of unlabeled data by considering different distributions of the labeled and unlabeled data and showing the migrating effect for semi-supervised learning. Extensive experiments have been performed in the context of image retrieval application. Our approach evaluates the value of unlabeled data from a new aspect and is aimed to provide a guideline on how unlabeled data should be used.

1. Introduction

Recently, there has been increasing interest in using unlabeled data for classification [1-8]. The motivation for this comes from the fact that labeled data is typically much harder to obtain compared to unlabeled data. This is valid in many applications, including web search, text classification, genetic research, and machine vision where an enormous amount of unlabeled data is available with little cost.

There are two existing approaches of taking advantage of unlabeled data. The first one is *semi-supervised learning* [1-4] and the second one is *active learning* [5-8]. In semi-supervised learning, one trains a classifier based on the labeled data as well as unlabeled data. Typically, a coarse classifier is first trained on the smaller labeled data set, and then it is used to give probabilistic labels to the unlabeled data. Finally, the enlarged, or hybrid data set consisting of both labeled and unlabeled data with probabilistic labeling is applied to re-train the classifier. In active learning, the coarse classifier is still based on the labeled data set, but instead of having all the unlabeled data labeled by the coarse classifier, a set of "most-informative" unlabeled data is selected. This set is then labeled by a human, as is the case of the relevance feedback approach of content-based image retrieval (CBIR) [7, 8]. The added small set of unlabeled data is believed to greatly enhance the construction of the new classifier. The advantage of the active learning is that as *little* data as possible will be labeled to achieve the improved performance.

There have been many studies on both active learning [5-8] and semi-supervised learning [1-4]. Past theoretical and experimental work showed that using the maximum-likelihood (ML) estimation approach (via EM or other numerical algorithms when unlabeled data was present) improved classification accuracy as more unlabeled data was added [2, 4]. Overall, these publications advance an optimistic view that unlabeled data can be profitably used wherever available.

However, in [2, 3], there are also reports that unlabeled data degrades the performances when it is added, e.g., Hughes phenomenon in [3]. Recently, Cozman et al. [9] conducted experiments on synthetic data aimed at understanding the value of unlabeled data. They reported that the classification accuracy could degrade more and more as more unlabeled data is added. Cozman et al. found that the reason for the degradation is the mismatch of the model assumption and the ground truth data distribution.

Considering all these aspects, several questions arise: when will unlabeled data help, and more importantly, how much do they help in classification and what are the underlying characteristics of the model that determines the usefulness of the unlabeled data? Conclusion from previous work [9, 1-4] on model assumption is that the ML estimator is unbiased and both labeled and unlabeled data contribute to a reduction in classification error by reducing variance as long as modeling assumptions match the ground-truth data. If model assumption does not match the ground-truth data, unlabeled data can improve or degrade the classification performance, depending on the complexity of the classifier compared with the size of the labeled training set [9, 4].

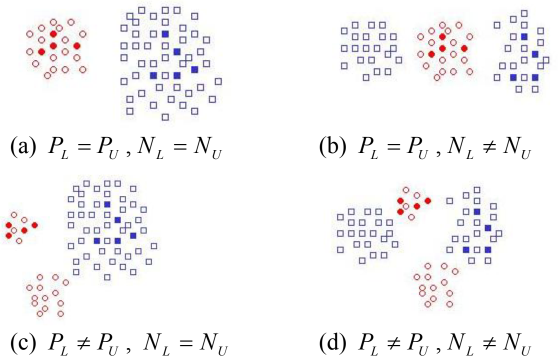
However, there is an underlying assumption that is often implicitly ignored or not addressed in the previous work [9, 4]: labeled data and unlabeled data are from the same distribution. [9, 4] focused on the probabilistic structure for the model and the ground-truth data and whether they match or not. In this paper, instead of looking into the model assumption issue, we investigate the value of unlabeled data in the semi-supervised learning when the labeled data and unlabeled data have different distributions.

Since this work is motivated from our research on content-based image retrieval, where a large amount of unlabeled images are available without cost, it is therefore natural for us to provide experiments on image classification problems for CBIR. The goal is to propose a guideline for any system that wishes to utilize unlabeled data to assist supervised learning in CBIR.

2. Problem Formulation

In image retrieval, there is a limited labeled training sample through relevance feedback [7, 8]. Limited training data would only result in weak classification. Considering that there are a large number of unlabeled images in a given database, we may use them to boost the weak classifier learned from the limited labeled data, since unlabeled data may contain useful information about the joint distributions over features. In such case, the hybrid training data set D consists of a labeled data set $L = \{(\mathrm{x}_i, c_i), i = 1,\ldots,N\}$ and an unlabeled data set $U = \{\mathrm{x}_i, i = 1,\ldots,M\}$. x_i is the feature vector of an image and c_i is its label that is either relevant or irrelevant, i.e., $c \in \{+1,-1\}$. The image retrieval system acts as a classifier to divide the images in the database into two classes, either relevant or irrelevant. The labeled data L is obtained by the query and relevance feedback, and the rest of the images in the database contribute to the unlabeled data set U. Denote the labeled positive images, i.e., relevant images, as P_L, and the labeled negative images, i.e., irrelevant images, as N_L and denote the unlabeled positive images as P_U and the unlabeled negative examples as N_U. Note that $L = P_L \cup N_L$, $U = P_U \cup N_U$, and $D = L \cup U$.

The labeled-unlabeled data problem is a combination of both *supervised* and *unsupervised* problems [10]. We want to build a classifier that generates a label $\hat{c}(\mathrm{x})$ for a given input image **x**. The classifier is built from a combination of existing labeled and unlabeled data sets. To build a classifier, we usually assume a model for the hybrid data (labeled data, or unlabeled data) under investigation. Some common assumptions used in the image retrieval community are, e.g., the feature independence among feature components, and mixture of Gaussians (or Gaussian) as the probabilistic structure of the image data set. When the assumed probabilistic structure matches the structure that generates the data, we say that the model structure is “correct”.



(a) $P_L = P_U$, $N_L = N_U$

(b) $P_L = P_U$, $N_L \neq N_U$

(c) $P_L \neq P_U$, $N_L = N_U$

(d) $P_L \neq P_U$, $N_L \neq N_U$

Figure 1: Four cases of distribution of *L* and *U*
(● : P_L , ○ : P_U , ■ : N_L , □ : N_U)

In the previous work [9, 4], the underlying assumption is that L and U are from the same distribution. In this work, we translate this assumption as $P_L = P_U$ and $N_L = N_U$ for simplicity. The “=” means that the labeled and unlabeled positive images are from the same distribution not that two sets are same. In most applications (e.g. CBIR) the size of U is much larger than the size of L. Therefore, when L and U are from the same distribution, it is equivalent to say L is like a subset of U. Then it is easy to understand the role of unlabeled data in semi-supervised learning, since adding more unlabeled data with the same distribution simply enlarge the representative training data set. When L and U are not from the same distribution, there are three cases (i) $P_L = P_U$ and $N_L \neq N_U$ (ii) $P_L \neq P_U$ and $N_L = N_U$ (iii) $P_L \neq P_U$ and $N_L \neq N_U$. An illustration of the above four cases is shown in Figure 1.

Previous work studied the value of unlabeled data under the correct and incorrect model assumption given L and U are from the same distribution. In this work, we take a new aspect to investigate the value of unlabeled data in the semi-supervised learning when L and U are not from the same distribution. Model selection is not our concern in this sense.

In the scenario of image retrieval, it is most likely to model the positive and negative images as two-class or $(1+x)$ -class classification problem. In this model, the positive images are from one class and the negative images are from one class or from multiple classes with unknown C. Although subclasses within a group can also exist for positive images, e.g., red car and white car within car category, they are usually treated as one class for simplicity. This corresponds to the cases (a) and (b) in Figure 1. Cases (c) and (d) correspond to $(y+1)$ -class and $(y+x)$-class classification problem and are considered as rare cases in image retrieval. In Section 3, we are investigating the value of unlabeled data for the cases (a) and (b) and leave the investigation of cases (c) and (d) in our future work.

3. Experiments

3.1. Case a: *L* and *U* with same distribution

In this first experiment, we test the value of unlabeled data under the same distribution (case a) on classification error. To visualize the effect of the unlabeled data, we adopt the same *LU-graph* as in [9].

Synthetic data is simulated to provide a better control of the data distribution. Gaussian model is used for both positive and negative samples. The centroid of the positive samples is placed at the origin while the negative samples have their centroid randomly placed with a distance of 0.5 to the origin. We assume feature independence and the variances of the positive and negative are randomly selected in the range (0.1~0.3). The feature dimension of each sample is 10. The labeled data set has a fixed size of 10. We vary the number of unlabeled data in semi-supervised learning from 0%, 20% to 80% and the size of the whole dataset from 50 to 800.

The expectation-maximization (EM) [10] algorithm is employed for semi-supervised learning. The Gaussian assumption is made for *L* and *U* in the EM algorithm. Figure 2 shows the precision in the top 20 retrieved positive samples. Precision measures the purity of the retrieved set, i.e., the percentage of relevant objects among those retrieved.

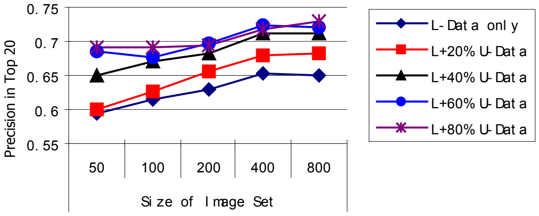


Figure 2: Precision for top 20 retrieved positive samples

It is clear in Fig. 2 that when *L* and *U* are from the same distribution, adding more unlabeled data certainly helps improve the precision. As more and more unlabeled data is added, the improvement becomes less significant. This agrees with the results by the other researchers.

3.2. Case b: *L* and *U* with different distribution

However if the labeled negative samples are not representative, i.e., $N_L \neq N_U$, adding more unlabeled data will degrade the classification performance as shown in Figure 3. In this case, positive samples are still placed at the origin and two negative classes have centroids placed with a distance of 0.5 to the origin. The labeled negative samples are from one negative class only, so they are not representative for the whole unlabeled data.

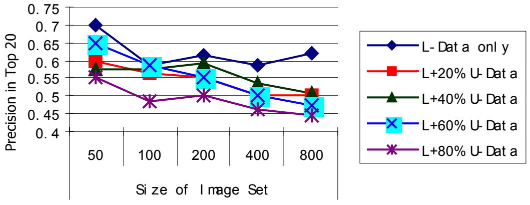


Figure 3: Precision for top 20 retrieved positive samples

It is clear that more unlabeled data will degrade the classification performance when *L* and *U* have different distributions. This is reasonable because semi-supervised learning assumes that the negative samples are generated from the same class. This assumption is no longer valid in our test. More unlabeled data just won't help.

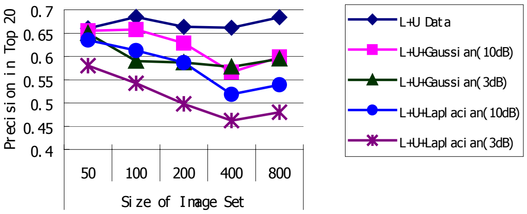


Figure 4: Precision for top 20 retrieved positive samples

3.3. Migrating effect of different distribution

The labeled and unlabeled data can contribute to a reduction in variance in semi-supervised learning under the ML estimation [4]. In this experiment, we test the value of unlabeled data by changing its probabilistic distribution structure with addition of different noises. In this work, we consider the case of adding noise to the unlabeled negative samples only so that $P_L = P_U$ but for the negative data it is varying from $N_L = N_U$ to $N_L \neq N_U$. *Gaussian* and *Laplacian* distributed noise [11] is considered. The purpose of this experiment is to see the migration effect of the unlabeled data for *L* and *U* from same to different distribution.

Figure 4 shows the results. The signal-to-noise (SNR) ratio is used to control the magnitude of the noise added to the unlabeled negative samples. In general, when the distribution of the unlabeled data is migrating from same distribution to different distribution, i.e., from $N_L = N_U$ to $N_L \neq N_U$, adding more unlabeled data (the size of *L* is fixed at 10) will degrade the precision. It is also interesting to point out that adding noise of different distribution (e.g., Laplacian) other than Gaussian, will have the worse effect on the value of unlabeled data. This is simply because adding noise other than Gaussian will change the distribution of the unlabeled data more noticeably.

3.4. Detect different distribution

The results in Section 3.1 to 3.3 show that unlabeled data will help if *L* and *U* are from the same distribution, otherwise, more unlabeled data will degrade the performance depending on the degree of difference between distribution of *L* and *U*, i.e., migrating effect. Then it is natural to ask a question: how to detect whether *L* and *U* are from the same distribution. The purpose is to use unlabeled data wisely only when it helps.

In statistics, parametric and non-parametric methods can be applied to test if two series of random variables are from the same distribution. Since it is difficult to have the accurate model of the image database, parametric testing is hard to apply. We will apply non-parametric testing instead. Commonly used non-parametric testing methods are *Dixon Test, Wilcoson Test*, and *Median Test* [12].

A subset of Corel dataset consisting of 14 categories with 99 images in each category is used for the experiments. Each time two categories are randomly chosen as positive and negative images. 20 images are randomly chosen as the labeled dataset while the rest used as unlabeled data. Table 1 shows the rejection value under different confidence level using Dixon test [12]. The image features are 37-dimension of color moments (9) in HSV color space, wavelet moments for texture (10), and edge-map based structure features (18) [13]. The distribution of unlabeled negative images is changed by the additive Laplacian noise. The larger rejection value is, the higher probability that *L* and *U* are from different distributions. The rejection value, e.g., 10 for 95% confidence level can be used as an indicator to measure the difference of distribution of *L* and *U*.

4. Discussion

Different from the previous work focusing on the probabilistic model selection, we take a new aspect on the value of unlabeled data in semi-supervised learning. We investigate the migrating effect of different probabilistic distribution of labeled and unlabeled data for image retrieval problem. The purpose is to provide a guideline to use unlabeled data wisely only when it helps. Our results show that unlabeled data helps only if *L* and *U* are from the same distribution. Investigation of cases (c) and (d) in Figure 1 will be our future work.

In this paper, we investigate the value of unlabeled data in the semi-supervised learning rather than in the active learning. Active learning [4] or selective sampling [5], studies the strategy for the learner, e.g., machine to actively select samples to query the teacher, e.g., user for labels, in order to achieve the maximal information gain in decision-making. How to select a most-informative subset of unlabeled data for active learning will be investigated in our future work as well.

Table 1: Detect different distribution of *L* and *U*

Rejection Value		$N_L : N_U$	$N_L : N_U + Lap.$ (3dB)	$N_L : N_U + Lap.$ (10dB)	$N_L : N_U + Lap.$ (20dB)	$N_L : N_U + Lap.$ (40dB)
Confidence Level	99%	12.1	16.1	15.2	13.3	13.2
	95%	8.8	11.9	11.6	10.9	10.3
	90%	8.1	10.3	10.4	8.9	8.5
	80%	7.1	8.6	9.1	8	7.7
	50%	4.9	6.1	6.4	6.1	5.7

5. References

[1] A. Blum, and T. Mitchell, "Combining labeled and unlabeled data with co-training," *Annual Conf. Comput. Learning Theory*, pp. 92-100, New York: ACM, 1998.

[2] K. Nigam, A McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Machine Learning*, 39, pp.1-32, 2000.

[3] 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 Trans. Geoscience and Remote Sensing*, vol. 32, no. 5, pp. 1087-1095, 1994.

[4] I. Cohen, N. Sebe, F. G. Cozman, M. C. Cirelo, and T. S. Huang, "Semi-supervised learning of classifiers: theory and algorithms for Bayesian network classifiers and applications to human-computer interaction," submitted to *IEEE Trans. on PAMI*, in revision, 2003.

[5] D. Cohn, Z. Ghahramani, and M. Jordan, "Active learning with mixture models," *Multiple Model Approaches to Modeling and Control*, R. Murray-Smith, T. Johanson (Eds.), pp. 167-183, 1997.

[6] Y. Freund, H. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," *Machine Learning*, 28, pp. 133-168, 1997.

[7] S. Tong, and E. Chang, "Support vector machine active learning for image retrieval," *Proc. of ACM Int'l. Conf. Multimedia*, pp. 107-118, 2001.

[8] I. J. Cox, *et al*. "The Bayesian image retrieval system, PicHunter: theory, implementation, and psychophysical experiments," *IEEE Trans. Image Proc.*, 9(1):20-37, 2000.

[9] F. G. Cozman, and I. Cohen, "Unlabeled data can degrade classification performance of generative classifiers," *Int'l Florida Artificial Intell. Society Conf.*, pp. 327-331, 2002.

[10] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, second edition, John Wiley & Sons, Inc., 2001.

[11] H. Stark, J. Woods, *Probability, Random Process and Estimation Theory for Engineers*, Prentice Hall, 1994.

[12] H. Motulsky, *Intuitive biostatistics: choosing a statistical test*, Oxford Univ. Press Inc., 1995.

[13] Q. Tian, *Content-based image visualization and retrieval for image libraries*, Ph.D. dissertation, University of Illinois at Urbana-Champaign, IL, U.S.A., 2002.