Combining Active Learning and Semi-supervised Learning Using Local and Global Consistency

Yingjie Gu[1,2], Zhong Jin[1], and Steve C. Chiu[2]

[1] Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
csyjgu@gmail.com, zhongjin@njust.edu.cn
[2] Department of Electrical Engineering, Idaho State University, Pocatello 83209-8060, USA
chiustev@isu.edu

Abstract. Semi-supervised learning and active learning are important techniques to solve the shortage of labeled examples. In this paper, a novel active learning algorithm combining semi-supervised Learning with Local and Global Consistency (LLGC) is proposed. It selects the example that can minimize the estimated expected classification risk for labeling. Then, a better classifier can be trained with labeled data and unlabeled data using LLGC. The experiments on two datasets demonstrate the effectiveness of the proposed algorithm.

Keywords: Active learning, semi-supervised learning, image classification.

1 Introduction

In traditional machine learning approaches to classification, only labeled examples are used to train the classifier. But in many real-world applications, there is a large number of unlabeled examples. Whereas labeled examples are usually difficult and expensive to obtain. Two typical methods to address this problem are semi-supervised learning [1] and active learning [2]. Semi-supervised learning combines both labeled examples and unlabeled examples to train a better classifier. Active learning usually selects a set of unlabeled instances for experts labeling, a better classifier can be trained by labeled examples afterwards.

The kernel of active learning is how to measure examples' value and which examples should be selected for labeling. There are many criteria in active learning to instruct examples selection. Uncertainty sampling is one of the most widely used criterion that queries the examples whose labels are most uncertain under the current classifier. Other criteria like variance reduction [3], Expected Model Change [4], Expected Error Reduction [5][6], and diversity [7] have also been widely applied to active learning.

With the same number of labeled examples, both active learning and semi-supervised learning usually perform better than supervised learning. It may make sense to utilize active learning in conjunction with semi-supervised learning.

C.K. Loo et al. (Eds.): ICONIP 2014, Part I, LNCS 8834, pp. 215–222, 2014.

Specifically, we firstly select a set of unlabeled examples to be labeled by experts. Then, both labeled examples and unlabeled examples are used to train classifiers. In [5], Zhu et al. combined active learning and semi-supervised learning using Gaussian Fields and Harmonic Functions (GFHF). Active learning is performed on top of the semi-supervised learning scheme by selecting examples to minimize the estimated expected classification risk.

Since Learning with Local and Global Consistency (LLGC) [8] presents a promising performance in semi-supervised learning, we explore the combination of active learning and LLGC in this paper. In active learning process, the example which can minimize the estimated expected classification risk is selected to be labeled. Then, a classifier is learned by LLGC with labeled data and unlabeled data. The experiments of image classification on two datasets demonstrate the effectiveness of the proposed algorithm.

The rest of this paper is organized as follows: In Section 2, we review semi-supervised Learning with Local and Global Consistency. The combination of active learning and LLGC is introduced in Section 3. In Section 4, we present the experimental settings and results. Finally, the conclusion and future work are discussed in Section 4.

2 Semi-supervised Learning with Local and Global Consistency

We begin by briefly describing the semi-supervised learning method LLGC [8]. Suppose there are l labeled examples $(x_1, y_1), ..., (x_l, y_l)$ and u unlabeled examples $x_{l+1}, ..., x_{l+u}$; usually $l \ll u$. y_i is the label of example x_i. For a c-class classification problem, $y_i \in \{1, 2, ..., c\}, i = 1, ..., l$. The labeled set and unlabeled set are denoted by $\mathcal{L}$ and $\mathcal{U}$, and $n = l + u$. The goal is to predict the labels of the unlabeled examples.

Let $\mathcal{F}$ denote the set of $n \times c$ matrices with nonnegative entries. Define a $n \times c$ matrix $Y \in \mathcal{F}$ with $Y_{ij} = 1$ if x_i is labeled as $y_i = j$ and $Y_{ij} = 0$ otherwise. A matrix $F \in \mathcal{F}$ is a matrix that labels all examples x_i with a label $y_i = argmax_{j \leq c} F_{ij}$. If F is defined as $F = [F_1^T, ..., F_n^T]^T$, F can be understander as a vectorial function which assigns a vector F_i to each example x_i. The LLGC algorithm is as follows:

1. Constrcut the affinity matrix W defined by $W_{ij} = exp(-||x_i - x_j||^2/2\sigma^2)$ if $i \neq j$ and $W_{ij} = 0$ if $i = j$.
2. Compute $S = D^{-1/2}WD^{-1/2}$ where D is a diagonal matrix with $D_{ii} = \sum_{j=1}^{n} W_{ij}$.
3. Iterate $F(t+1) = \sigma SF(t) + (1-\sigma)Y$ until convergence, where σ is a parameter in $(0, 1)$.
4. Define $F^* = \lim_{t \to \infty} F(t)$. The label of x_i is predicted as $y_i = \arg\max_{j \leq c} F^*_{ij}$.

We firstly construct a graph $G = (V, E)$ on $\mathcal{L} \cup \mathcal{U}$, where the vertex set V is the set of all examples and the edges E are weighted by W. Then, the weight

matrix W is normalized symmetrically. In the iteration, each examples receives information from its neighbors (first term), and retains its initial information (second term). The information is spread symmetrically since S is a symmetric matrix. Finally, the label of each unlabeled examples is predicted as the class of which it has received most information during the iteration process.

By computing the limit of the sequence $\{F(t)\}$, we can obtain

$$F^* = (1-\alpha)(I-\alpha S)^{-1}Y \tag{1}$$

for classification, which is equivalent to

$$F^* = QY \tag{2}$$

where $Q = (I-\alpha S)^{-1}$. Since S is fixed, Q is also fixed in the learning process.

A regularization framework was also proposed by Zhou et al. for this method. The cost function associated with F with regularization parameter $\mu > 0$ is defined as

$$\mathcal{Q}(F) = \frac{1}{2}\left(\sum_{i,j=1}^{n} W_{ij}\left\|\frac{1}{\sqrt{D_{ii}}}F_i - \frac{1}{\sqrt{D_{jj}}}F_j\right\|^2 + \mu\sum_{i=1}^{n}\|F_i - Y_i\|^2\right) \tag{3}$$

The optimal decision function is $F^* = \arg\min_{F\in\mathcal{F}} \mathcal{Q}(F)$. More on this semi-supervised learning framework can be found in [8].

3 Active Learning

In this section, we propose to perform active learning with LLGC. The basic idea of the proposed active learning is to select the example that can minimize the classification risk of the examples.

With both labeled examples and unlabeled examples, we can train a classifier (decision function F) using LLGC. The class of unlabeled example x_i is predicted as $y_i = \arg\max_{j\le c} F^*_{ij}$. Suppose $P(y_i|x_i)$ is the probability distribution of the examples' labels. We assume that the distribution $P(y_i|x_i)$ can be estimated based on decision function F.

$$P(y_i = j|x_i) = \frac{F_{ij}}{\sum_{t=1}^{c} F_{it}} \tag{4}$$

We define the true risk $\mathcal{R}(P)$ of the classification based on labels' distribution P. Thus

$$\mathcal{R}(P) = \sum_{i=1}^{n}(1 - \max_{j=1,...,c} P(y_i = j|x_i)) \tag{5}$$

If we perform active learning to select an unlabeled example x_k for experts labeling, we will receive an answer y^*_k ($y^*_k \in \{1,...,c\}$). Before we selecting x_k for labeling, $Y_{kj} = 0$ ($j = 1,...,c$). After labeling x_k and adding (x_k, y^*_k) to labeled set, the matrix Y should be updated and denoted by $Y^{+(x_k,y^*_k)}$ where

$Y^{+(x_k,y_k^*)}_{k,y_k^*} = 1$. The decision function F and the probability distribution P will also change

$$F^{+(x_k,y_k^*)} = QY^{+(x_k,y_k^*)} \tag{6}$$

$$P^{+(x_k,y_k^*)}(y_i = j|x_i) = \frac{F_{ij}^{+(x_k,y_k^*)}}{\sum_{t=1}^{c} F_{it}^{+(x_k,y_k^*)}} \tag{7}$$

If (x_k, y_k^*) is added to the labeled set, the estimated classification risk is

$$\mathcal{R}(P^{+(x_k,y_k^*)}) = \sum_{i=1}^{n}(1 - \max_{j=1,\ldots,c} P^{+(x_k,y_k^*)}(y_i = j|x_i)) \tag{8}$$

Before we querying experts about the label of x_k, the true label y_k^* is unknown. But we can obtain the labels' distribution $P(y_i|x_i)$ from decision function F. Therefore, the expected classification risk after querying x_k is estimated as

$$\mathcal{R}(P^{+x_k}) = \sum_{j=1}^{c} P(y_k = j|x_k)\mathcal{R}(P^{+(x_k,j)}) \tag{9}$$

We aim to select the example that can minimize the expected estimated risk. Therefore, the index of the selected example is

$$s = \arg\min_{k\in\{l+1,\ldots,n\}} \mathcal{R}(P^{+x_k}) \tag{10}$$

Once the label y_s^* of the example x_s is queried from experts, (x_s, y_s^*) will be added to the labeled set. The label matrix Y will be updated to $Y^{+(x_s,y_s^*)}$ and the decision function will be retrained by equation (6). In fact, the update operation of label matrix Y is only to change one element in Y, namely set Y_{s,y_s^*} to be 1. The retraining step $F^{+(x_s,y_s^*)} = QY^{+(x_s,y_s^*)}$ is equivalent to update the y_s^*-th column of the matrix F.

$$F^{+(x_s,y_s^*)}_{\cdot y_s^*} = F_{\cdot y_s^*} + Q_{\cdot y_s^*} \tag{11}$$

where $F_{\cdot y_s^*}$ and $Q_{\cdot y_s^*}$ denote the y_s^*-th column of matrices F and Q. Of course $F^{+(x_s,y_s^*)}_{\cdot j} = F_{\cdot j}$ if $j \neq y_s^*$. It is easy to prove that the equation (6) is equivalent to equation (11). But the computation of equation (11) is much faster than equation (6).

The process of the proposed active learning combining LLGC is concluded in Table 1. It is the procedure of selecting one example for experts labeling. In applications, the examples selection often repeats many times until the stop criterion is reached.

4 Experiment

In order to assess the effectiveness of the proposed technique, we evaluate and compare five active learning methods:

Table 1. The process of the proposed active learning algorithm

Input:

Initial labeled data set $(x_1, y_1), ..., (x_l, y_l)$, unlabeled data set $x_{l+1}, ..., x_{l+u}$, the guassian kernel parameter σ, the tradeoff parameter α

Output:

The selected example

Procedure:

Construct label matrix Y, compute weight matrix W and S, Q, F

For $k = l + 1 : n$

For $y_k = 1 : c$

$F^{+(x_k,y_k)} = F$, $F^{+(x_k,y_k)}_{\cdot y_k} = F_{\cdot y_k} + Q_{\cdot y_k}$

$P^{+(x_k,y_k)}(y_i = j|x_i) = \frac{F^{+(x_k,y_k)}_{ij}}{\sum_{t=1}^{c} F^{+(x_k,y_k)}_{it}}$

$\mathcal{R}(P^{+(x_k,y_k)}) = \sum_{i=1}^{n}(1 - \max_{j=1,...,c} P^{+(x_k,y_k)}(y_i = j|x_i))$

End

$\mathcal{R}(P^{+x_k}) = \sum_{j=1}^{c} P(y_k = j|x_k)\mathcal{R}(P^{+(x_k,j)})$

End

$s = \arg\min_{k \in \{l+1,...,n\}} \mathcal{R}(P^{+x_k})$

Return x_s

- Random Sampling with LLGC classifier (RS+LLGC), which randomly selects examples for labeling and uses LLGC classifier.
- Most Uncertain with LLGC classifier (MU+LLGC), which selects the most uncertain example from LLGC classifier for labeling. The index of the most uncertain example is

$$s = \arg\min_{i=l+1,...,n} F_{ij_1} - F_{ij_2} \tag{12}$$

  where $j_1 = \arg\max_{j=1,...,c} F_{ij}$, $j_2 = \arg\max_{j=1,...,c, j \neq j_1} F_{ij}$.
- Multiclass-level uncertainty with SVM classifier (MCLU+SVM), which was proposed in [9].
- MinRisk+GFHF, which was proposed in [5].
- MinRisk+LLGC, which is proposed in this paper. The parameter α is set to 0.99 and σ is set to 0.1.

In the following sections, we carry out classification experiments on two real-world data sets to compare different active learning algorithms quantitatively.

4.1 Handwritten Digits Recognition

The USPS handwritten digits data set is used in this experiment. The data set contains 8-bit gray-scale images of '0' through '9'. The size of each image is 16×16 pixels. Thus, each digit image is represented as a 256-dimensional vector.

On this data set, we used digits 1, 2, 3, and 4 in our experiments as the four classes. 500 examples from each class are randomly selected so there are

totally 2000(500×4) examples. Only 1 example from each class is randomly selected as initial labeled example. Thus there are 4 labeled examples and 1996 unlabeled examples. We apply each active learning algorithm to select k ($k = 1, 2, ..., 10$) examples for labeling. A classifier can be trained with LLGC or SVM method. Lastly, we predict the labels of the rest unlabeled examples and compute the classification accuracy. The experiments are repeated for 30 times and the average accuracy is obtained.

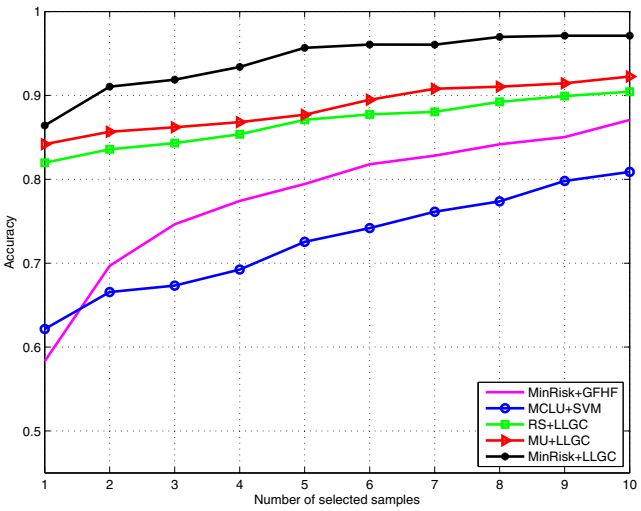


Fig. 1. The average classification accuracy on usps dataset

Fig.1 shows the average classification accuracy versus the number of examples selected by active learning methods. As can be seen, our MinRisk+LLGC algorithm significantly outperforms the other active learning algorithms. MU+LLGC performs the second best. The active learning combining GFHF (MinRisk+ GFHF) is not better than our proposed method. MCLU+SVM is worse than others since it is unable to use unlabeled examples to train a classifier.

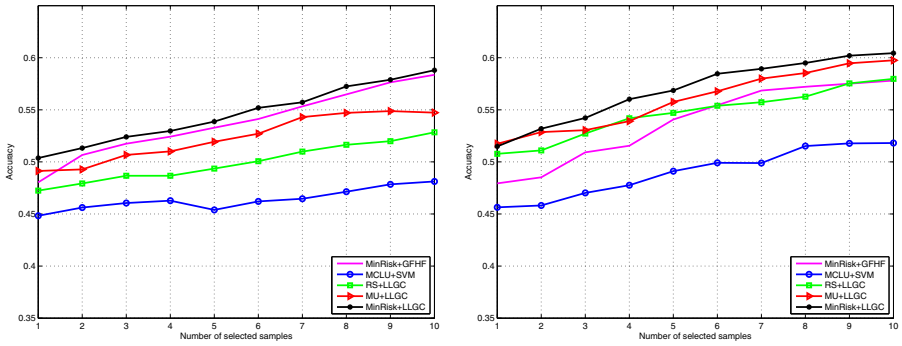
4.2 Terrain Classification

In this section, we apply active learning algorithms to terrain classification problems. Terrain image dataset used in the experiment was constructed by us from the Outex Database [10], which consists of two data sets: Outex-0 and Outex-1. Each of them includes 20 outdoor scene images and the size of each image is 2272×1704. The images are marked as one type of bush, grass, tree, sky, road, and building. The marked area of each image is cut into patches with size 64×64 and each patch is regarded as an example. Two examples of each class are shown

in Fig. 2. Both color histogram feature and LBP feature are extracted and combined to represent each example. We extract 100 patches of each class (totally 600 patches) to construct a pool of unlabeled data set for examples selection. Firstly, only 1 example of each class is labeled as initial labeled set. Then, active learning is used to select k ($k = 1, 2, ..., 10$) examples for labeling. Lastly, a classifier is trained and the labels of the unlabeled examples are predicted.

Fig. 2. Examples of Outex from categories: sky, tree, bush, grass, road, and building

The average classification accuracies on Outex-0 and Outex-1 are shown in Fig.3. As can been seen, our MinRisk+LLGC outperforms the other algorithms in most of the cases. MinRisk+GFHF performs the second best on Outex-0 while worse than MU+LLGC on Outex-1. MCLU+SVM performs the worst on two datasets since it is a supervised learning method that does not use unlabeled data in learning.



(a) The classification accuracy on Outex-0 (b) The classification accuracy on Outex-1

Fig. 3. The average classification accuracy on Outex-0 and Outex-1

To sum up, semi-supervised learning (LLGC, GFHF) performs better than supervised learning (SVM) with the same number labeled examples. Our proposed MinRisk+LLGC outperforms MinRisk+GFHF, MU+LLGC, and RS+LLGC in most of the cases.

5 Conclusion

In this paper, a novel active learning algorithm which combining semi-supervised learning with LLGC is proposed. The example that can minimize the estimated expected classification error is selected for labeling. Experiments on two datasets indicate that the proposed algorithm can be highly effective.

MinRisk+LLGC is a single-mode active learning algorithm that selects only one example each time. In the future, we will expend this method into a batch-mode active learning.

Acknowledgements. This work is partially supported by National Natural Science Foundation of China under Grant Nos. 61373063, 61233011, 61125305, 61375007, 61220301, and by National Basic Research Program of China under Grant No. 2014CB349303.

References

1. Zhu, X.: Semi-supervised learning literature survey. Computer Science 2, 3 (2006)
2. Settles, B.: Active learning literature survey. University of Wisconsin, Madison (2010)
3. Ji, M., Han, J.: A variance minimization criterion to active learning on graphs. In: International Conference on Artificial Intelligence and Statistics, pp. 556–564 (2012)
4. Settles, B., Craven, M., Ray, S.: Multiple-instance active learning. In: Advances in Neural Information Processing Systems, NIPS, pp. 1289–1296. MIT Press (2008)
5. Zhu, X., Lafferty, J., Ghahramani, Z.: Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In: ICML 2003 Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining, pp. 58–65 (2003)
6. Guo, Y., Greiner, R.: Optimistic active-learning using mutual information. In: IJCAI, vol. 7, pp. 823–829 (2007)
7. Chakraborty, S., Balasubramanian, V., Panchanathan, S.: Dynamic batch mode active learning. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2649–2656. IEEE (2011)
8. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: NIPS, vol. 16, pp. 321–328 (2003)
9. Demir, B., Persello, C., Bruzzone, L.: Batch-mode active-learning methods for the interactive classification of remote sensing images. IEEE Transactions on Geoscience and Remote Sensing 49(3), 1014–1031 (2011)
10. University of oulu texture database, `http://www.outex.oulu.fi/temp/`