Semi-supervised Classification with Active Query Selection

Jiao Wang and Siwei Luo

School of Computer and Information Technology, Beijing Jiaotong University,
Beijing 100044, China
wangjiao0828@163.com

Abstract. Labeled samples are crucial in semi-supervised classification, but which samples should we choose to be the labeled samples? In other words, which samples, if labeled, would provide the most information? We propose a method to solve this problem. First, we give each unlabeled examples an initial class label using unsupervised learning. Then, by maximizing the mutual information, we choose the samples with most information to be user-specified labeled samples. After that, we run semi-supervised algorithm with the user-specified labeled samples to get the final classification. Experimental results on synthetic data show that our algorithm can get a satisfying classification results with active query selection.

1 Introduction

Recently, there has been great interest in Semi-supervised classification. The goal of semi-supervised learning is to use unlabeled data to improve the performance of standard supervised learning algorithms. Since in many fields, obtaining labeled data is hard or expensive, semi-supervised learning methods with small labeled sample size is of great use.

In case the unsupervised learning methods can separate the points well (see e.g. Fig.1a), there is no need for semi-supervised methods. However, in case of noise (see e.g. Fig.1b), or in case of two modes which belong to two different classes overlap (see e.g. Fig.1c), semi-supervised learning with a few labeled points in each class can improve the performance significantly.

A number of algorithms have been proposed for semi-supervised learning, including EM [8], Co-training [1, 14], Tri-training [15], random field models [9, 12], graph based approaches [2, 6, 13]. Different methods have different assumptions, and can be used in different situation. Especially, when data resides on a low-dimensional manifold within a high-dimensional representation space, semi-supervised learning methods should be adjusted to work on manifold. Belkin gives a solution to this problem with manifold Regularization methods in [4].

Query selection is extensively studied in the supervised framework. In [10], the queries are selected to minimize the version space size for support vector machine. In [7], a committee of classifiers is employed, and a point is queried whenever the committee members disagree. Many other methods are proposed to actively choose the samples in supervised learning, but few are done to choose samples in semi-supervised learning.

The labeled samples play an important role in semi-supervised learning. Then a question rises: which samples should be the labeled samples? Among the existing semi-supervised learning methods, some choose the labeled samples manually[6], to do this, one has to have some domain knowledge of which samples need most to be labeled; some choose the labeled samples randomly, which may not contain the “right” samples; and in [11], Zhu et al. choose the samples actively by greedily selecting queries from the unlabeled data to minimize the estimated expected classification error, but Zhu’s active learning method can only be used together with his semi-supervised learning method.

In this paper, we give a more general and automatic query selection method in the semi-supervised framework. Our method can be applied to most of the existing semi-supervised learning methods. It is the pre-process of the existing semi-supervised methods. The main idea is to consider which samples, if labeled, would give more information. Following this idea, we use the mutual information $I(Y; y_*)$ (y_* represents one sample’s class label and Y represents the whole sample’s class labels) as a measure of active query selection. By maximizing the mutual information, we get the sample which needs most to be labeled. Using this method, we can choose the samples to be labeled actively and automatically, and it does not need any domain knowledge.

In this paper, in order to explain our method, we work with the Laplacian Eigenmaps [3] and manifold Regularization [4] of Belkin to show the entire process. We can see how the active query selection method works on manifold. We do not claim that this method can only use on manifold, and indeed we aim to illuminate that applying our method, to any semi-supervised method, would always yield satisfying results.

This paper is organized as follows: In section 2, we introduce our algorithm in a brief way. Section 3 gives details of every part of our algorithm. Experimental results on synthetic data are shown in section 4, followed by conclusions in section 5.

Fig. 1. (a) Example of situation in which unsupervised learning methods (here we use Laplacian Eigenmaps) can work well. (b)(c) Examples of situations in which unsupervised learning can not give satisfying results, and they need some labeled samples to help.

2 Our Algorithm

To explain the entire process of our active query selection method, we work with the Laplacian Eigenmaps [3] and manifold Regularization [4] of Belkin. The steps are as follow:

- Step 1. Give each sample (unlabeled) an initial class label using unsupervised learning. Here, we use Laplacian Eigenmaps to map the sample to a real value function f .
- Step 2. By maximizing the mutual information $I(Y; y_*)$ (y_* represents one sample's class label and Y represents the whole sample's class labels), we actively choose the samples with most uncertain class labels to be user-specified labeled samples.
- Step 3. Give the chosen samples their class label.
- Step 4. Run the semi-supervised algorithm (here, we use the manifold Regularization algorithm) with the user-specified labeled samples to get the final classification.

3 Details of Our Method

3.1 Using Laplacian Eigenmaps to Get Initial Class Label

Given a sample set $x_1, \dots, x_n \in R^m$. Construct its neighborhood graph $G = (V, E)$, whose vertices are sample points $V = \{x_1, \dots, x_n\}$, and whose edge weights $\{w_{ij}\}_{i,j=1}^n$ represent appropriate pairwise similarity relationships between samples. For example, w_{ij} can be the radial basis function:

$$w_{ij} = \exp\left(-\frac{1}{\sigma^2} \sum_{d=1}^m (x_{id} - x_{jd})^2\right) \quad (1)$$

where σ is a scale parameter. The radial basis function of w_{ij} ensure that nearby points are assigned large edge weights.

We first consider two-class situation. Assume that f is a real value function whose value is bounded from 0 to 1 (0 and 1 each represents a class label). $y_i = f(x_i)$, $Y = (y_1, y_2, \dots, y_n)^T$. Laplacian Eigenmaps try to minimize the following objective function

$$\sum_{ij} (y_i - y_j)^2 w_{ij} \quad (2)$$

By minimizing this objective function, we get $y_1, y_2, \dots, y_n$, the initial class label of each sample, with $y_i \in [0, 1]$.

3.2 Using Mutual Information to Choose the Samples with Most Uncertain Class Labels

This is our active query selection step. And we use the mutual information $I(Y; y_*)$ (y_* represents one sample's class label and Y represents the whole sample's class labels) as a measure of query selection. By maximizing the mutual

information, we get the sample which would give most information, that is, which needs most to be labeled.

In order to calculate $I(Y; y_*)$, inspired by the work of [5], we define a Gaussian random field on the vertices of V

$$p(y) \propto \exp\{-\lambda y^T \Delta y / 2\} \quad (3)$$

where $\Delta = D - W$, and D is a diagonal matrix given by $D_{ii} = \sum_{j=1}^n W_{ij}$.

The mutual information between Y and y_* is the expected decrease in entropy of Y when y_* is observed:

$$\begin{aligned} I(Y; y_*) &= H(Y) - E\{H(Y | y_*)\} \\ &= (1/2) \log(1 + p(y_*)(1 - p(y_*))x_*^T H^{-1} x_*) \end{aligned} \quad (4)$$

where $H = \nabla^2(-\log p(Y))$, and ∇^2 is the Hessian matrix.

The best sample to label is the one that maximizes $I(Y; y_*)$. And the mutual information is largest when $p(y_*) \approx 0.5$, i.e., for samples with most information.

3.3 Using the User-Specified Labeled Samples to Get the Final Classification

After we actively choose the sample to give label, we can run the semi-supervised classification methods to get the final result. In Manifold Regularization methods of Belkin, the author minimizes the following cost function

$$\min_{f \in H} H[f] = \frac{1}{l} \sum_{i=1}^l (f(x_i) - y_i)^2 + \gamma_A \|f\|_K^2 + \gamma_l \sum_{i,j=1}^n (f(x_i) - f(x_j))^2 W_{ij} \quad (5)$$

Where l is the number of labeled samples. γ_A, γ_l are regularization parameters.

$\|f\|_K^2$ is some form of constraint to ensure the smoothness of the learned manifold.

Here, the l samples in the above cost function are not chosen randomly or manually as in the original work of Belkin. But rather, they are chosen with the active query selection methods discussed in 3.2.

4 Experimental Results

As we point out at the beginning of this paper, unsupervised learning can not work well in case of noise, and in case of two modes which belong to two different classes overlapped. In these situations, semi-supervised learning with a few labeled samples can help.

Using some synthetic data, we show that our active query selection method can choose the most informative samples to give labels. Fig.2 (a) is a noise case of fig.1 (a), and without labeled samples, the Laplacian Eigenmaps can not find a satisfying

classification (the yellow curve). Using our active query selection method, the algorithm chooses some samples to be labeled, these samples are shown in (b) with purple color. After that, user give the class label of these chosen samples (the red and blue samples in (c)), each color represents a class), then, with these user-specified labeled samples, manifold regularization method find the more satisfying classification as shown in (c) (the yellow curve).

Fig. 2. (a) Laplacian Eigenmaps can not find a satisfying classification without labeled samples. (b) The samples automatically chosen to be labeled. (c) The manifold regularization results with the labeled samples.

5 Conclusions

A key problem of semi-supervised learning is to choose the most informative samples to be labeled at the very beginning of semi-supervised algorithms. Using mutual information, we give a solution to this problem. Our method of samples chosen can apply to most of the existing semi-supervised learning methods, and in this paper, we combine it with manifold regularization to show how it works.

We also do experiments on some synthetic data, and yield satisfying results. In future works, we will try this method on some real world experiments. Another problem of semi-supervised learning is how many labeled sample are suitable, for example, should we choose five samples to give label, or, should we choose ten? In future work, we will consider this problem in the framework of the active query selection of this paper.

Acknowledgements

The research is supported by national natural science foundations of china (60373029), the Research Fund for the Doctoral Program of Higher Education of China (20050004001) and Co-Construction Project of Key Subject of Beijing.

References

1. A. Blum and T. Mitchell: Combining labeled and unlabeled data with co-training. In Proceedings of the 11th Annual Conference on Computational Learning Theory. Madison, WI, pp.92–100, (1998).
2. A. Blum, S. Chawla: Learning from Labeled and Unlabeled Data using Graph Mincuts. ICML (2001).

3. Belkin M., Niyogi P: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, June (2003)
4. Belkin M., Niyogi P., Sindhwani V: On Manifold Regularization. Department of Computer Science, University of Chicago, TR-2004-05.
5. B. Krishnapuram, D. Williams, Ya Xue, A. Hartemink, L. Carin, and M. A. T. Figueiredo: On Semi-Supervised Classification. *NIPS* (2004).
6. D. Zhou, O. Bousquet, T.N. Lal, J. Weston and B. Schoelkopf: Learning with Local and Global Consistency. *NIPS* (2003).
7. Freund, Y., Seung, H. S., Shamir, E., & Tishby, N.: Selective sampling using the query by committee algorithm. *Machine Learning*, 28, 133-168. (1997).
8. K. Nigam: Using Unlabeled Data to Improve Text Classification. PhD thesis, Carnegie Mellon University Computer Science Dept, (2001).
9. M. Szummer and T. Jaakkola: Partially labeled classification with markov random walks. *NIPS* (2001).
10. Tong, S., and Koller, D.: Support vector machine active learning with applications to text classification. *ICML* (2000).
11. Xiaojin Zhu, J. Lafferty, and Z. Ghahramani: Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions. *ICML* (2003).
12. Xiaojin Zhu, Z. Ghahramani, and J. Lafferty: Semi-supervised learning using Gaussian fields and harmonic functions. *ICML* (2003).
13. Xiaojin Zhu: Semi-Supervised Learning with Graphs. PhD thesis, Carnegie Mellon University Computer Science Dept, (2005).
14. Zhou, Z.-H., & Li, M.: Semi-supervised regression with co-training. *International Joint Conference on Artificial Intelligence*, (2005).
15. Zhou, Z.-H., & Li, M.: Tri-training: exploiting unlabeled data using three classifiers. *IEEE Trans. Knowledge and Data Engineering*, 17, 1529–1541, (2005).