Locally Non-negative Linear Structure Learning for Interactive Image Retrieval

Lei Bao[1,2], Juan Cao[1], Tian Xia[1], Yong-Dong Zhang[1], Jintao Li[1]
[1]Laboratory for Advanced Computing Technology Research, ICT, CAS, Beijing 100190, China
[2]Graduate University of Chinese Academy of Sciences, Beijing 100049, China
{baolei, caojuan, txia, zhyd, jtli}@ict.ac.cn

ABSTRACT
A successful interactive image retrieval system is expected to quickly return as many relevant results as possible while costing less users' effort. Considering these system demands, firstly we propose a novel semi-supervised learning algorithm called Locally Non-negative Linear Structure Learning (LNLS), which is based on the assumption that the labels of each data should be sufficiently smooth with respect to the locally non-negative linear structure of dataset. It has two main merits: first, it is robust to the small sample learning problem since it learns structure from both labeled and unlabeled data; second, by emphasizing the non-negativity of locally linear structure, this algorithm preserves the non-negative inherent characteristic of image data and can truly reveal the intrinsic structure of the images corpus, especially the asymmetric relationship between images. Meanwhile, we explore an online updating algorithm for LNLS to tackle the large computation cost. Thus the model can be generalized to the new queries or the newly-labeled samples without retraining. Furthermore, an active learning method for LNLS is proposed to make the most of users' effort to improve the learner. The encouraging experimental results demonstrate the effectiveness and efficiency of our proposed methods.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *search process, relevance feedback.*

General Terms: Algorithms, Experimentation

Keywords: Locally non-negative linear structure, interactive image retrieval, active learning

1. INTRODUCTION
With the rapid growth of digital images and the success of new participatory web technologies, interactive image retrieval has attracted much attention in recent years.

Users always expect to be quickly presented with a large number of relevant results while expending as little effort as possible [1]. Since that, to design a successful interactive system, we must balance these demands across two interrelated system components: the learning/search strategy and the interaction strategy [1]. That means the former should be flexible to the small labeled samples with low time complexity, and the latter must be painless for users while providing the most information to the learner.

Considering the above constraints in interactive retrieval systems, we propose a novel semi-supervised learning algorithm called Locally Non-negative Linear Structure Learning (LNLS), which is based on the assumption that the labels of each data should be sufficiently smooth with respect to the locally non-negative linear structure of dataset. Firstly, the LNLS trains the model from both labeled and unlabeled data, and can achieve a desirable performance even if the size of labeled samples is small. Secondly, the LNLS focuses on the locally linear structure of dataset which represents each data as a liner combination of its neighbors. The LNLS is superior to the popular graph-based methods [2, 3] as the latter are sensitive to the parameter of similarity metric, and can not reflect the asymmetric relationship between images. Finally, we claim that the non-negativity as an inherent characteristic of image data should be preserved in locally linear structure, where each image is reasonably represented as an additive combination of its neighbors rather than a subtractive combination, and the label score as well. Since the LNLS conforms to the image's physical property, a desirable performance is expectable.

Another contribution in this paper is that, an online updating algorithm is proposed for LNLS to tackle the large computation cost. When dealing with new queries or increasing labeled samples, the model can adaptively update without retraining. It is feasible to be applied in real-time system.

Finally, to make the most of users' effort, we propose an active learning method for LNLS, which actively asks users to label the most informative samples so that the retrieval performance could be improved most efficiently.

The rest of this paper is organized as follows: Section 2 details the LNLS and its online updating algorithm. An active learning algorithm is provided in Section 3, followed by experimental results in Section 4. Finally, the paper is concluded in Section 5.

2. LOCALLY NON-NEGATIVE LINEAR STRUCTURE LEARNING

2.1 Motivation
A principled approach to semi-supervised learning is to design a classifying function which is sufficiently smooth with respect to the intrinsic structure revealed by known labeled and unlabeled points [3]. Graph-based methods describe this structure as a graph, in which vertices denote labeled and unlabeled samples and edges reflect the similarities between samples, and assume label scores

MM'09, October 19–24, 2009, Beijing, China.

smoothness over the graph. Based on this assumption, many popular graph-based methods are proposed [2, 3], and have been widely applied in image and video content analysis [4, 5, 10]. However, the graph-based methods are sensitive to similarity metric. What's more, in image corpus, the symmetric similarity can not reflect the real asymmetric relationship between images. Given two images "sky" and "airplane in sky", whatever which one is chosen as relevant, the other's score is equal based on graph-based methods. Actually, since "sky" co-occurs with "airplane in sky", when the former is set as relevant the latter's score should be higher than the opposite case that the latter is set as relevant. As a result, the graph-based learning methods in image retrieval sometimes lead to an unsatisfactory performance.

Obviously, besides graph structure, the locally linear structure, which represents each data as a liner combination of its neighbors, can also reveal the intrinsic structure of dataset and should be preserved in low-dimensional space. This is demonstrated by the previous work Locally Linear Embedding (LLE) [6]. Inspired by LLE, [7] proposed Linear Neighborhood Propagation (LNP) on the assumption that the label of samples should be smooth with respect to the locally linear structure. Firstly, the locally linear structure is not sensitive to similarity metric. Secondly, since relationship score from data A to B represents the contribution of A to reconstruct B, this reveals the asymmetric relationship between images. Furthermore, considering the non-negativity of image data (actually the non-negativity is a very popular property existing in most data), we also believe that the non-negativity should be preserved in locally linear structure, where each image data is reasonably represented as an additive combination of its neighbors rather than a subtractive combination, and the label score as well. Finally, the Locally Non-negative Linear Structure (LNLS) learning is proposed on this assumption, and it is desirable to outperform the previous work.

2.2 Formulation

Given a dataset $\chi = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_l, \mathbf{x}_{l+1}, \cdots, \mathbf{x}_n\} \subset R^d$, the first l data $\mathbf{x}_i (1 \le i \le l)$ are labeled as $y_i \in \{-1, 1\} (1 \le i \le l)$ and the remaining $\mathbf{x}_i (l+1 \le i \le n)$ are unlabeled. The smooth assumption on locally non-negative linear structure can be formulized by two steps.

Step 1. Learning the locally non-negative linear structure of χ.

The basic assumption of locally linear structure of dataset is that each data can be linearly reconstructed from its neighbors. This process can be defined as the reconstruction error

$$E(\mathbf{W}) = \left\| \mathbf{X} - \mathbf{W}\mathbf{X} \right\|^2, \ s.t. \ \mathbf{W}\mathbf{1} = \mathbf{1}, \tag{1}$$

where $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_l, \mathbf{x}_{l+1}, \cdots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times d}$, $\mathbf{W} \in \mathbb{R}^{n \times n}$ and $\mathbf{1}$ is an n-dimensional vector with all elements are 1. w_{ij} is the reconstruction weights from $\mathbf{x}_j$ to $\mathbf{x}_i$ when $\mathbf{x}_j$ is a k-nearest-neighbor of $\mathbf{x}_i$ by some distance metric, $\mathbf{x}_j \in N(\mathbf{x}_i)$. $w_{ij} = 0$ when $\mathbf{x}_j \notin N(\mathbf{x}_i)$. What's more, considering the non-negativity of image data, we further constrain $w_{ij} \ge 0$, which ensures the image data can be reasonably represented as an additive combination of its neighbors rather than a subtractive combination. What's more, $\mathbf{W}$ is asymmetric. It is obvious the contribution from image "sky" to reconstruct image "airplane in sky" is more than that from the latter to the former. This asymmetry of $\mathbf{W}$ is enforced by its non-negativity which leads to part-based representation [8]. As a result, the LNLS preserves the non-negativity of image data and reflects the asymmetric relationship between images.

Step 2. Predicting the label score according to the non-negative reconstruction weights $\mathbf{W}$.

Based on the smooth assumption, the prediction process is defined as follows:

$$E(\mathbf{f}) = \left\| \mathbf{f} - \mathbf{W} \right\|^2 \mathbf{f} + (\mathbf{f} - \mathbf{y})^T \mathbf{C} (\mathbf{f} - \mathbf{y}), \tag{2}$$

where $\mathbf{f} = [f_1, f_2, \cdots, f_n]^T \in \mathbb{R}^n$ indicates the predicted label score of χ and $\mathbf{C} \in \mathbb{R}^{n \times n}$ is a diagonal matrix satisfied: $c_i = C_l > 0$ when $1 \le i \le l$; $c_i = C_u \ge 0$ when $l+1 \le i \le n$. C_l and C_u are parameters for the soft constraints on labeled and unlabeled samples. The first term stands for the smoothness constraints, which means the label score f_i of $\mathbf{x}_i$ should match the reconstructed score from its neighbors' as much as possible. The second term denotes the fitting constraints, which means the label scores $\mathbf{f}$ should not change too much from the given label scores $\mathbf{y}$. It is noted from (2), with the constraints $\mathbf{W}\mathbf{1} = \mathbf{1}$ and $w_{ij} \ge 0$ in step 1, the reconstructed score is a convex combination of its neighbors', which ensure the f_i of unlabeled $\mathbf{x}_i$ will not be out of the range [-1, 1]. Under this condition, it is reasonable to regard the reconstructed score as a predicted label score. That is another benefit from the non-negative reconstruction weights.

2.3 Solution

The first optimization problem in step 1 can be solved by non-negative matrix factorization, as it can be rewritten as follows:

$$\mathbf{X} = \mathbf{W}\mathbf{X}, \ s.t. \ \mathbf{W} \ge 0, \mathbf{W}\mathbf{1} = \mathbf{1}. \tag{3}$$

Resorting to the auxiliary function [8], we derive the following iterative bound optimization algorithm.

$$w_{ik}^{(t)} = \frac{\left[\mathbf{X}\mathbf{X}^T\right]_{ik} + \alpha}{\left[\mathbf{W}^{(t-1)}\mathbf{X}\mathbf{X}^T\right]_{ik} + \alpha\left[\mathbf{W}^{(t-1)}\mathbf{1}\right]_{ik}} w_{ik}^{(t-1)}, \tag{4}$$

where α is a positive constant weighting the constraint $\mathbf{W}\mathbf{1} = \mathbf{1}$. Due to the space limit, we omit the proofs for the convergence.

For the second optimization problem in (2), the following solution can be easily obtained by differentiating $E(\mathbf{f})$ with respect to $\mathbf{f}$

$$\mathbf{f}^* = (\mathbf{M} + \mathbf{C})^{-1} \mathbf{C}\mathbf{y}, \tag{5}$$

where $\mathbf{M} = (\mathbf{I} - \mathbf{W})^T (\mathbf{I} - \mathbf{W})$.

2.4 Online Updating Algorithm

Generally, given a corpus, the computing W in step 1 is offline. Since $\mathbf{W}$ is sparse especially in non-negativity condition and the process of computing $\mathbf{W}$ is iterative, step 1 is flexible to large-scale corpus. However, step 2 should be executed online in interactive image retrieval system. When dealing with new queries or newly-labeled samples, we have to recompute $(\mathbf{M} + \mathbf{C})^{-1}$. It is time-consuming with computational cost $o(n^3)$. Actually, by some matrix methods, the inversion can be updated from the last round result, and the computational cost can be reduced to $o(n^2)$.

Given the results of last round: $\mathbf{f}^{(t-1)}$ and $\mathbf{S}^{(t-1)} = \left(\mathbf{M} + \mathbf{C}^{(t-1)}\right)^{-1}$, when $(\mathbf{x}_k, y_k)$ is the latest labeled sample, the new $\mathbf{S}^{(t)}$ becomes

$$\mathbf{S}^{(t)} = \left(\mathbf{M} + \mathbf{C}^{(t)}\right)^{-1} = \left(\mathbf{M} + \mathbf{C}^{(t-1)} + \left(C_l - C_u\right)\mathbf{e}_k\mathbf{e}_k^T\right)^{-1}. \quad (6)$$

Based on Sherman-Morrision-Woodbery formula, after some calculating, we can derive

$$\mathbf{S}^{(t)} = \mathbf{S}^{(t-1)} - \frac{\left(C_l - C_u\right)}{1 + \left(C_l - C_u\right)\mathbf{S}_{kk}^{(t-1)}}\mathbf{S}_{\cdot k}^{(t-1)}\mathbf{S}_{k\cdot}^{(t-1)}, \quad (7)$$

$$\mathbf{f}^{(t)} = \mathbf{f}^{(t-1)} + \frac{C_l y_k - \left(C_l - C_u\right) f_k^{(t-1)}}{1 + \left(C_l - C_u\right)\mathbf{S}_{kk}^{(t-1)}}\mathbf{S}_{\cdot k}^{(t-1)}. \quad (8)$$

Based on the above formulas, when $\mathbf{S}^0 = \left(\mathbf{M} + C_u\right)^{-1}$ is calculated offline, in the online interactive retrieval process, the model can be updated in $o(n^2)$ computational cost without retraining.

3. ACTIVE LEARNING

In order to alleviate the effort required for users, the active learning is proposed to ask the user to label the most informative unlabeled sample such that the retrieval performance could be improved most efficiently [1]. The key issues are how to define the most informative samples and how to choose them efficiently.

Here, we define the most informative sample for the proposed LNLS as the one which minimizes the Bayesian classification error. In LNLS learning, the probability of $\mathbf{x}_i$ labeled as 1 can be estimated as $\left(1 + f_i\right)/2$. Then before a sample $\mathbf{x}_k$ is selected, the Bayesian classification error can be calculated as.

$$\begin{aligned}\varepsilon\left(\mathbf{f}\right) &= \sum_{i=1}^{n}\left([\mathrm{sgn}\left(f_i\right) \neq 1]\frac{\left(1 + f_i\right)}{2} + [\mathrm{sgn}\left(f_i\right) \neq -1]\frac{\left(1 - f_i\right)}{2}\right), \\ &= \frac{1}{2}\sum_{i=1}^{n}\left(1 - \left|f_i\right|\right)\end{aligned} \quad (9)$$

when $\mathbf{x}_k$ is selected, y_k is obtained from users, we have a new label score vector $\mathbf{f}^{+(\mathbf{x}_k, y_k)}$ and the new Bayesian classification error. Although we don't know what answer we will receive, we can calculate the expected error:

$$\varepsilon\left(\mathbf{f}^{+\mathbf{x}_k}\right) = \frac{\left(1 + f_i\right)}{2}\varepsilon\left(\mathbf{f}^{+(\mathbf{x}_k, y_k = 1)}\right) + \frac{\left(1 - f_i\right)}{2}\varepsilon\left(\mathbf{f}^{+(\mathbf{x}_k, y_k = 0)}\right). \quad (10)$$

We choose the most informative sample $\mathbf{x}_{k^*}$ that minimizes the expected Bayesian classification error:

$$k^* = \arg\min_k \varepsilon\left(\mathbf{f}^{+\mathbf{x}_k}\right). \quad (11)$$

According to the online updating algorithm in subsection 2.4, we can compute $\mathbf{f}^{+(\mathbf{x}_k, y_k)}$ by (8) without retraining.

4. EXPERIMENTAL RESULTS

In the experiments, we use the 2000 images from COREL as our database, which are categorized into 20 groups. Each group consists of 100 images and represents one semantic topic. For feature representation of images, we extract 64-dimentional HSV color histogram and 75-dimensional edge distribution histogram for relative performance comparison.

To evaluate the performances of the proposed algorithms for interactive image retrieval, we conduct experiments on two aspects: learning strategy and interaction strategy. Average precision (AP) and mean average precision (MAP) are adopted as evaluation metric, in which the relevance judgments are based on whether the query image and the retrieval image belong to the same group [10].

4.1 Learning Strategy

We compare the proposed Locally Non-negative Linear Structure learning (LNLS) with Support Vector Machine (SVM) [10], Gaussian Random Field (GRF) [2], Local and Global Consistency (LGC) [3], and Linear Neighborhood Propagation (LNP) [7].

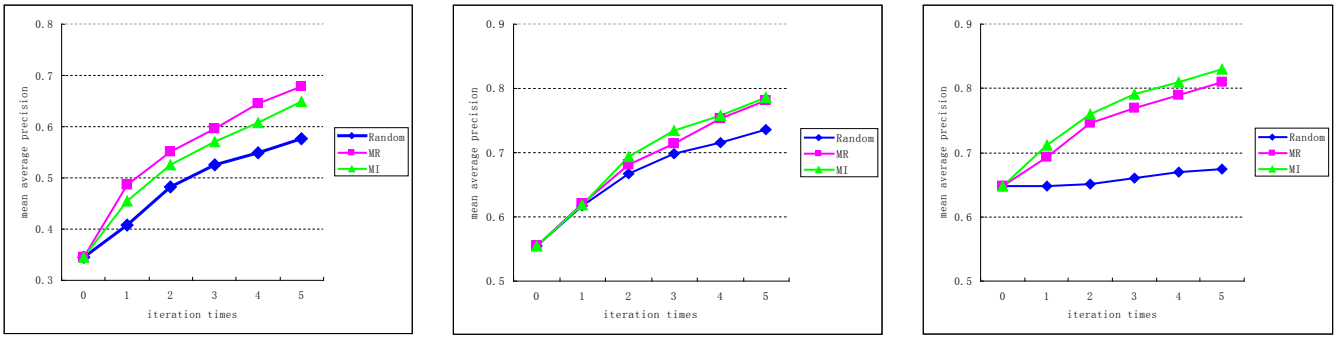
The penalty C and scaling parameter δ of SVM is set to 100 and 0.1. For GRF and LGC, we build a weighted K-mutual nearest neighbor graph, where K is fixed at 50. L1 is chosen to be the distance metric and the weights (similarities) are calculated using Laplace kernel with parameter 0.05 as recommended in [5]. The parameter for LGC also is fixed at 0.99 consistent with [2]. For LLP and LNLS, we also use the same distance metric and K as GRF and LGC. The parameters C_l and C_u are fixed to 0.1 and 0.01, respectively. The parameter α of LNLS is fixed at 0.15.

In learning strategy, the learning algorithms have to face two conditions. The one is initial retrieval stage, where there is only one query sample available. The other is relevance feedback retrieval stage, where a few labeled samples are available. To simulate the first condition, we randomly pick an image as query. To simulate the second condition, we conduct experiments on varied labeled size. Since the relevant samples are always hard to obtained, the percentage of the relevant images in labeled set is fixed to 10%. In each run, we calculate AP on the top 200 results. The presented MAP is based on 20 groups with each group running for 10 times.

The results of the five methods in the above two conditions are presented in Table 1, where SVM and GRF are excluded in the first condition since they can not work without irrelevance samples. From Table 1, we can find that, 1) all these four semi-supervised methods (GRF, LGC, LNP and LNLS) outperform the traditional supervised method SVM in all experiments, which indicates that semi-supervised methods are robust to the small sample learning problem, even in one query condition, since they utilize the unlabeled data; 2) all the locally linear structure based methods (LNP and LNLS) outperform the graph-based methods (GRF and LGC), since the latter is sensitive to similarity metric

Table 1. Performance of different learning methods

Label Size	SVM	GRF	LGC	LNP	LNLS
1	-----	-----	0.3210	0.3400	**0.3867**
10	0.2078	0.3626	0.3249	0.3376	**0.4007**
20	0.1960	0.4071	0.3394	0.4145	**0.4557**
30	0.2448	0.4497	0.3789	0.4641	**0.4986**
40	0.2924	0.4836	0.4242	0.5024	**0.5301**
50	0.3518	0.5112	0.4807	0.5380	**0.5738**



(a) initialized with 1 relevant image (b) initialized with 5 relevant images (c) initialized with 10 relevant images

Figure 1. Performance comparison with different interaction strategies

and can not reveal the asymmetric relationship between images; 3) LNLS consistently outperforms the other three semi-supervised methods, which proves the effectiveness and efficiency of LNLS, since it preserves the non-negativity of image data and well reveals the asymmetric relationship between images.

4.2 Interaction Strategy

For interaction strategy, we compare the proposed active learning for LNLS ,which aims to choose the most informative samples (MI), with the randomly selecting scheme (Radom) and the most relevant scheme (MR), which is recommended by [4] since the lack of relevant samples in image retrieval.

To evaluate the performance of these three schemes, we conduct them initialed with various relevant sizes: 1, 5, and 10, and feedback 5 times with 10 images returned in each iteration. Figure 1 illustrates these results. We can see that: 1) the most informative scheme and the most relevant scheme outperform the randomly selecting scheme and have a stable improvement with the iteration time increasing, which demonstrates these interaction strategies actually do some work to maximize the user's efforts; 2) when initialized with only 1 relevant sample, the most relevance scheme outperform the proposed most informative one due to the lack of relevance samples as shown in Figure 1(a); however, with the increasing of the initialized label size, the advantage is no longer obvious as shown in Figure 1(b); when the initialized relevance sample size increases to 10, the latter surpasses the former as shown in Figure 1(c), which demonstrates that the most relevant scheme and the most informative scheme have their own favorite scenario, respectively. This conclusion will help us to choose interaction strategy for interactive retrieval system.

5. CONCLUSION

A successful interactive image retrieval system is expected to quickly return as many relevant results as possible while costing less users' effort. Taking into account these system demands, we proposed a novel semi-supervise learning algorithm called Locally Non-negative Linear Structure Learning (LNLS), and its corresponding online updating algorithm and active learning algorithm. Firstly, LNLS preserves the non-negativity of image data and reveals the asymmetric relationship between images, which ensures LNLS outperforms the previous semi-supervised methods with significant improvement. Secondly, by an online updating algorithm, LNLS can be generalized to the new queries or the newly-labeled samples without retraining, which makes LNLS suitable to a real-time system. Finally, an active learning algorithm for LNLS is proposed to make most of users' effort to improve performance. Experiments on both learning strategy and interaction strategy demonstrate the effectiveness and efficiency of these proposed methods in interactive image retrieval.

6. ACKNOWLEDGMENTS

This work was supported by National Basic Research Program of China (973 Program, 2007CB311100), National High Technology and Research Development Program of China (863 Program, 2007AA01Z416), National Nature Science Foundation of China (60873165, 60802028).

7. REFERENCES

[1] Thomas S. Huang, et al. Active learning for interactive multimedia retrieval. In *Proc. of the IEEE,* vol. 96, no. 4, 648-667, 2008.

[2] X. Zhu, et al. Semi-supervised learning using gaussian fields and harmonic functions. In *Proc. of Int'l Conf. on Machine Learning,* 2003.

[3] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. SchÄolkopf. Learning with local and global consistency. In *Proc. of Advances of Neural Information Processing*, 2004.

[4] J.R. He, et al. Manifold-Ranking Based Image Retrieval. In. *Proc. of ACM Multimedia,* 2004.

[5] M. Wang, et al. Video annotation by graph-based learning with neighborhood similarity. In. *Proc. of ACM Multimedia,* 2008.

[6] S.t. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*: vol. 290, no. 5500, 2323-2326. 2000.

[7] F. Wang, et al. Semi-Supervised classification using linear neighborhood propagation. In. *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition,* 2006.

[8] D.D. Lee, H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*: vol. 401, 788-791. 1999

[9] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In. *Proc. of ACM Multimedia,* 2001

[10] J. Tang, et al. Structure-Sensitive Manifold Ranking for Video Concept Detection. In. *Proc. of ACM Multimedia,* 2007