Positive and Negative Label Propagations

Olga Zoidi, Anastasios Tefas, Nikos Nikolaidis, Senior Member, IEEE, and Ioannis Pitas, Fellow, IEEE

Abstract

This paper extends the state-of-the-art label propagation (LP) framework in the propagation of negative labels. More specifically, the state-of-the-art LP methods propagate information of the form "the sample i should be assigned the label k." The proposed method extends the state-of-the-art framework by considering additional information of the form "the sample i should not be assigned the label k." A theoretical analysis is presented in order to include negative LP in the problem formulation. Moreover, a method for selecting the negative labels in cases when they are not inherent from the data structure is presented. Furthermore, the incorporation of negative label information in two multigraph LP methods is presented. Finally, a discussion on the proposed algorithm extension to out of sample data, as well as scalability issues, is presented. Experimental results in various scenarios showed that the incorporation of negative label information increases, in all cases, the classification accuracy of the state of the art.

Keywords: Action recognition, face recognition, graph-based semisupervised learning (GSSL), label propagation (LP).

1 Introduction

Label Propagation (LP) is a commonly used method for classifying a set of partially labeled data by considering both the label information of the labeled data and the structure of both the labeled and unlabeled data. Most LP methods operate on similarity graphs [1]. In these methods, the graph nodes represent the visual data and the graph edge weights represent their pairwise similarities, which depend on the features that were selected for data representation. Then, label inference is performed along graph paths that connect labeled nodes to unlabeled ones.

The most widely used LP [1] performs LP with local and global consistency. It is essentially a manifold regularization method. For each label, one function is considered that assigns each graph node with a real value. The initialization of the function is performed by assigning the value 1 to the nodes that are known to have a certain label and 0 to the remaining nodes. The optimization framework then regulates the function values so that nodes with initial nonzero values maintain their original value and adjacent nodes with high weights are assigned similar values. The result of manifold regularization for a certain label on the graph nodes indicates the association of the nodes to this label. More specifically, the nodes with high function value have high association to the respective label. Finally, label assignment to the unlabeled samples is performed by selecting the label that corresponds to the function with the largest value for the respective sample node.

LP is a special case of transductive semisupervised learning. Transductive semisupervised learning refers to the construction of classifiers that exploit class information from a set of training (labeled) data,

along with the structure information of the test (unlabeled) data, in order to learn a local representation of the data space that spans on the available train and test data. As a result, such classifiers cannot be employed on "unknown" data that belong neither to the originally available training nor to test data. All LP methods, including transductive semisupervised classification methods, consider information from a few training samples with known class information and the structure of all data in the training and test data set. Even the imposition of additional discriminant constraints, in the form "manifold values on samples that belong to the same class should have small variance" and "manifold values on samples that belong to different classes should have large variance" in the optimization framework of such methods, is based exclusively on the class information of the training samples. However, there are certain applications in which additional information for the data can be exploited that cannot be incorporated in the existing frameworks, as described in the following paragraph.

Let $\mathcal{S} = \{(\mathbf{x_i}, l_i), i = 1, ..., N | \mathbf{x}_i \in R^M, l_i \in \mathcal{L}\}$ be a data set of N samples, each one belonging in one of the classes of $\mathcal{L}$, as shown in Fig. 1(a). The class (label) from a few samples (those with filled symbols) is known beforehand, while the class of the rest is unknown. We want to propagate the label information from the labeled data in $\mathcal{S}$ to the unlabeled ones. Let us assume that it is a *priori* known that the data set $\mathcal{S}$ was constructed from the union of two subsets of samples $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$, as shown in Fig. 1.(b), as follows: $\mathcal{S}_1 = \{(\mathbf{x_i}, l_i), i = 1, ..., N_1 | x_i \in \mathcal{R}^M, l_i \in \mathcal{L}_1\}$, $\mathcal{S}_2 = \{(\mathbf{x}_{N_1+i}, l_{N_1+i}), i = 1, ..., N_2 | \mathbf{x}_{N_1+i} \in R^M, l_{N_1+i} \in \mathcal{L}_2\}$, $N = N_1 + N_2$, $\mathcal{L} = \mathcal{L}_1 \cup \mathcal{L}_2$, where $\mathcal{L}_1 = \{L_1, L_2, L_3\}$ and $\mathcal{L}_2 = \{L_2, L_3, L_4\}$. We notice that the set $\mathcal{S}_1$ does not contain data that belong to class L_4 . Similarly, the set $\mathcal{S}_2$ does not contain data that belong to class L_1 . Therefore, in order to optimally classify the unlabeled data in the set $\mathcal{S}$ through LP, apart from the state-of-the-art LP assumptions, the employed framework should also consider the above mentioned observations. As a result, apart from the standard conditions, we have the following.

- The labels of the initial labeled data should be preserved.
- Data that are similar to each other should be assigned the same label.

The employed LP framework should be able to ensure satisfaction of the following additional conditions.

- Data that belong to the set $\mathcal{S}_1$ should not be assigned the label L_4 .
- Data that belong to the set $\mathcal{S}_2$ should not be assigned the label L_1 .

Fig. 1(f) shows that the application of the state-of-the-art LP in $\mathcal{S}$ does not take into advantage the last two conditions, therefore, it does not lead to optimal classification results. Moreover, the application of the state-of-the-art LP separately in subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ does not lead to optimal classification results either. Fig. 1(c) shows that the LP on $\mathcal{S}_2$ achieves perfect classification accuracy. However, this is not the case for the LP in $\mathcal{S}_2$, as shown in Fig. 1(d). More precisely, the samples in $\mathcal{S}_2$ belonging to classes L_2 and L_3 do not contain adequate information for the structure of the respective classes. Therefore,

LP performance is poor for these classes. The classification results for separate LP on $\mathcal{S}_1$ and $\mathcal{S}_2$ are summarized in Fig. 1(e).

In this report, a novel LP method is presented that tackles the general task of positive and negative label propagations (PNLPs). More specifically, the task of "positive" LP tries to solve the problem of spreading the label information from a small set of data with known label to a much larger set of data with unknown label. The word "positive" has been added in LP (though not existing in the literature) to distinguish between the classical ("positive") LP and the proposed PNLPs. The ("positive") LP algorithm assigns the same label to data that are considered similar, according to some similarity measure. The task of negative LP solves the dual problem, i.e., instead of propagating the information that the ith sample has the lth label, we propagate the information that the ith sample does not have the kth label. This means that, in negative LP, the actual label information of the data is not known. Since negative propagation propagates label restrictions for the data, it can be considered as label constraint propagation.

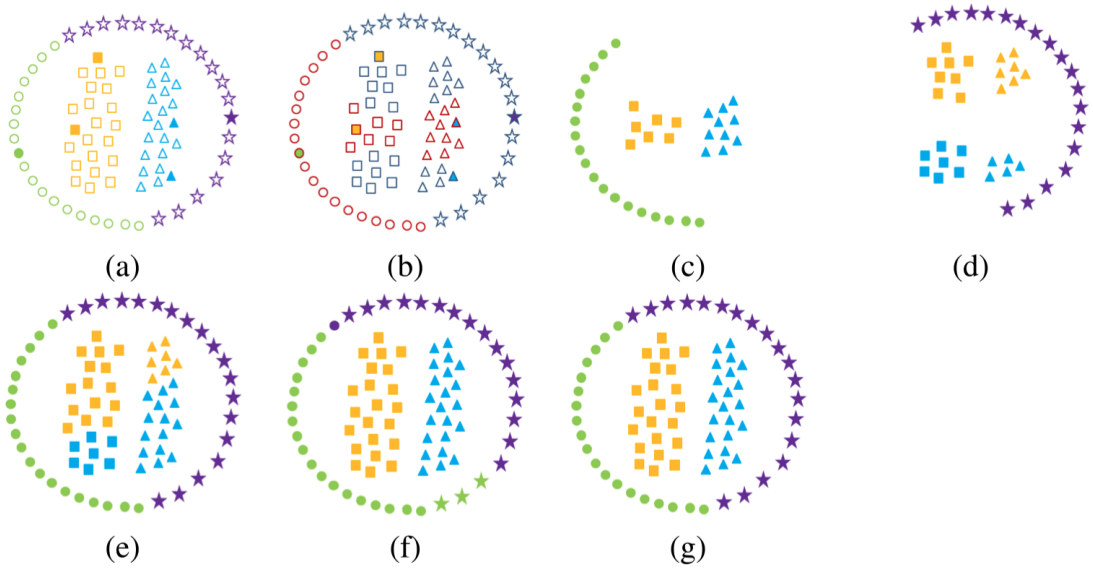

Figure 1: Fig. 1. (a) Data in $\mathcal{S}$ that belong to four classes. The data with label L_1 are denoted with a green circle. The data with label L_2 are denoted with a yellow square. The data with label L_3 are denoted with a blue triangle. The data with label L_4 are denoted with a purple star. The data with colored filling denote the initially labeled data. (b) Data in sets $\mathcal{S}_1$ and $\mathcal{S}_2$. The data in $\mathcal{S}_1$ are denoted with a red border. The data in $\mathcal{S}_2$ are denoted with a blue border. The data with colored filling denote the initially labeled data. (c) LP results on the data in $\mathcal{S}_1$. (d) LP results on the data in $\mathcal{S}_2$. (e) LP results on the data in $\mathcal{S}$ when it is performed separately on $\mathcal{S}_1$ and $\mathcal{S}_2$. (f) LP results on all data in $\mathcal{S}$. (g) PNLP results on all data in $\mathcal{S}$.

2 Related works

2.1 POSITIVE LABEL PROPAGATION

The task of positive LP tries to solve the problem of spreading the label information from a small set of data with known labels to a much larger set of data having unknown labels. Positive LP is simply called LP problem in the literature. It assigns the same label to data that are considered to be similar, according to some similarity measure. LP solves the following regularization framework,which

was introduced in [2].

Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^M$ be the set of N data that belong to classes $\mathcal{L} = \{1, ..., L\}$. We consider that each sample belongs only to one class. We consider that the class labels $l(\mathbf{x}_i) \in \mathcal{L}, i = 1, ..., N_l$ of N_l data are known. $\mathcal{G} = (\mathcal{X}, \mathcal{E})$ is the graph, whose edges are the data entries $\mathbf{x}_i$ in the set $\mathcal{X}$ and whose edges represent pairwise data relationships. A graph edge that connects nodes i and j is assigned with a value (similarity weight) W_{ij} that indicates the similarity between the two graph nodes. Usually, this similarity weight is computed according to the heat kernel equation [3]

$$W\left(\mathbf{x}_i, \mathbf{x}_j\right) = W_{ij} = \exp\left(-\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{\sigma}\right) \tag{1}$$

where σ is the mean edge length distance among neighbors. A function $\mathbf{F} : \mathcal{X} \rightarrow \mathbb{R}^{N \times L}$ is defined that assigns a vector of dimension L on each graph node. The vector elements represent one score value for each label. Finally, $\mathbf{Y} \in \mathbb{R}^{N \times L}$ is a matrix that represents the initial labels with entries

$$Y_{ij} = \begin{cases} 1, & \text{if } l\left(\mathbf{x}_i\right) = l_j \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

LP is performed by minimizing the regularization framework

$$\mathcal{Q}(\mathbf{F}) = \frac{1}{2}\operatorname{tr}\left(\mathbf{F}^T\mathbf{L}\mathbf{F}\right) + \frac{\mu}{2}\operatorname{tr}\left[(\mathbf{F}-\mathbf{Y})^T(\mathbf{F}-\mathbf{Y})\right] \tag{3}$$

where $\mu > 0$ is a regularization parameter, and $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$, where $\mathbf{D} = \operatorname{diag}_i\left\{\sum_{j=1:N} W_{ij}\right\}$ is the normalized graph Laplacian. The first term in (3) represents the clustering assumption, i.e., similar data are assigned the same label, while the second term ensures that the label of the initially labeled data remains unchanged. Minimization of $\mathcal{Q}(\mathbf{F})$ with respect to $\mathbf{F}$ leads to the following optimal solution for $\mathbf{F}^*$:

$$\mathbf{F}^* = \mu(\mathbf{L} + \mu\mathbf{I})^{-1}\mathbf{Y} \tag{4}$$

The definition of $\mathbf{Y}$ and the clustering assumption postulate that a high value of F^*_{ij} corresponds to a high probability that the ith sample is assigned the jth label. Therefore, label assignment for sample $\mathbf{x}_i$ is performed according to

$$l_i = \arg\max_j \{F_{ij}\} \tag{5}$$

2.2 NEGATIVE LABEL PROPAGATION

Negative LP refers to the dual problem of positive negative propagation, i.e., instead of propagating the information that the ith sample has the lth label, we propagate the labeling constraint that the ith sample does not have the kth label. This means that, in negative LP, the actual label information is

not known. This fact renders labeling constraints as less informative than positive labels.

LP is equivalent to negative propagation under the following formulation. Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be the set of N data that belong to classes $\mathcal{L} = \{l_1, \ldots, l_L\}$. As in last section, we consider that each sample belongs only to one class and that the class labels $l(\mathbf{x}_i) \in \mathcal{L}, i = 1, \ldots, N_l$ of N_l data are known. In terms of negative LP, this information is equivalent to the claim that the sample $\mathbf{x}_i, i = 1, \ldots, N_l$ does not have any of the labels in $\mathcal{L}^i = \mathcal{L} - \{l(\mathbf{x}_i)\}$. Let $\mathcal{G} = (\mathcal{X}, \mathcal{E})$ be the graph, whose nodes are the data entries $\mathbf{x}_i$ in the set $\mathcal{X}$ and whose edge weights are the pairwise data similarities W_{ij} according to the heat kernel equation (1). Let $\Psi \in \mathbb{R}^{N \times L}$ be the initial state matrix, with entries

$$\Psi_{ij} = \begin{cases} 1, & \text{if } l_j \in \mathcal{L}^i \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Negative LP is performed by minimizing the regularization problem defined by

$$\mathcal{Q}(\Phi) = \frac{1}{2} \operatorname{tr}\left(\Phi^T \mathbf{L} \Phi\right) + \frac{\mu}{2} \operatorname{tr}\left[(\Phi - \Psi)^T (\Phi - \Psi)\right] \tag{7}$$

where μ is a regularization parameter, $\mathbf{L}$ is the normalized graph Laplacian, and $\Phi \in \mathbb{R}^{N \times L}$ is a matrix that assigns a score on each sample for each label. Similarly to the case of positive LP, the optimal solution for Φ is given by

$$\Phi^* = \mu(\mathbf{L} + \mu \mathbf{I})^{-1} \Psi \tag{8}$$

The definition of Ψ and the clustering assumption postulates that a small value of Φ^*_{ij} indicates a high probability that the ith sample has the jth label. Therefore, label assignment for sample $\mathbf{x}_i$ is performed according to

$$l_i = \arg\min_j \left\{\Phi^*_{ij}\right\} \tag{9}$$

Moreover, by definition, $\Psi = \mathbf{1}_{N \times L} - \mathbf{Y}$, where $\mathbf{1}_{N \times L} \in \mathbb{R}^{N \times L}$ is a matrix of ones. By substituting Ψ in (13), we obtain

$$\Phi^* = \mu(\mathbf{L} + \mu \mathbf{I})^{-1} (\mathbf{1}_{N \times L} - \mathbf{Y}) \tag{10}$$

or by considering (8)

$$\Phi^* = \mu(\mathbf{L} + \mu \mathbf{I})^{-1} \mathbf{1}_{N \times L} - \mathbf{F}^* \tag{11}$$

The first term in (11) depends on the data graph, and it is constant, regardless of the label initialization. Moreover, from (11), we notice that Φ^* becomes minimum when $\mathbf{F}^*$ becomes maximum. Therefore, it can be concluded that

$$l_i = \arg\max_j \left\{F^*_{ij}\right\} = \arg\min_j \left\{\Phi^*_{ij}\right\}. \tag{12}$$

This means that the classification results when either the positive or the negative LP formulation is

employed are equivalent. From the above, it can be concluded that one positive label for some sample is equal to $L-1$ negative labels for the same sample, where L is the total number of labels. Only in the case of binary classification ($L=2$), positive and negative labels have equal strength. However, even though label constraints are less informative than positive labels, their incorporation in the LP framework will increase its overall informativeness, as will be discussed in next section.

2.3 POSITIVE AND NEGATIVE LABEL PROPAGATIONS

Let $\mathcal{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}\subset\mathbb{R}^M$ be the set of N data that belong to one of the classes in $\mathcal{L}=\{l_1,\ldots,l_L\}$. We assume that two kinds of information are known for some of the samples: positive labels and negative labels (labeling constraints). Positive label information $\mathcal{P}\subset\mathcal{X}\times\mathcal{L}$ is of the form

$$\mathcal{P}=\{(\mathbf{x}_i,l_i)\,,i\in\{p_1,\ldots,p_q\}\} \tag{13}$$

where the pair $(\mathbf{x}_i,l_i)$ denotes that the ith sample has the label l_i , while negative label information $\mathcal{N}\subset\mathcal{X}\times\mathcal{L}$ is of the form

$$\mathcal{N}=\{(\mathbf{x}_i,l_i')\,,i\in\{n_1,\ldots,n_m\}\} \tag{14}$$

where the pair $(\mathbf{x}_i,l_i')$ denotes that the ith sample does not have the label l_i . A novel algorithm is devised that propagates both kinds of information concurrently on all samples in $\mathcal{X}$. To this end, a graph $\mathcal{G}=(\mathcal{X},\mathcal{E})$ is constructed similar to the one in Section III. A classification function $\mathbf{F}:\mathcal{X}\rightarrow\mathbb{R}^L$ is defined on the graph nodes that assign a real value for each label. For each label, the function should assign similar values to nodes with high similarity to each other. High F_{ij} values indicate high probability that the ith sample has the jth label. Finally, two matrices $\mathbf{Y}^+$ and $\mathbf{Y}^-$ are defined to represent the positive and the negative label information, with entries

$$Y_{ij}^{+}=\begin{cases}1, & \text{if from prior knowledge } l\left(\mathbf{x}_i\right)=l_j\\ 0, & \text{otherwise}\end{cases} \tag{15}$$

$$Y_{ij}^{-}=\begin{cases}1, & \text{if from prior knowledge } l\left(\mathbf{x}_i\right)\neq l_j\\ 0, & \text{otherwise.}\end{cases} \tag{16}$$

By extending the regularization framework in (3), in order to incorporate the negative label information, the following objective function is defined:

$$\begin{aligned}\mathcal{Q}(\mathbf{F})=&\frac{1}{2}\operatorname{tr}\left(\mathbf{F}^T\mathbf{L}\mathbf{F}\right)+\frac{\mu}{2}\\ &\times\left[\mu_1\operatorname{tr}\left(\left(\mathbf{F}-\mathbf{Y}^{+}\right)^T\left(\mathbf{F}-\mathbf{Y}^{+}\right)\right)\right.\\ &\left.-\mu_2\operatorname{tr}\left(\left(\mathbf{F}-\mathbf{Y}^{-}\right)^T\left(\mathbf{F}-\mathbf{Y}^{-}\right)\right)\right].\end{aligned} \tag{17}$$

The first term in (17) is the graph regularization term. The second term forces the initially labeled

samples to retain their initial label. The third term restricts the initially negative labeled samples in obtaining the label indicated in $\mathcal{N}$. The parameter $0 < \mu < 1$ regulates the significance of the overall positive and the negative label information in the optimization framework. Moreover, parameters $0 < \mu_1 < 1$ and $0 < \mu_2 < 1$ regulate the relative significance between the positive and the negative label information. μ_1 and μ_2 are restricted so that $\mu_1 + \mu_2 = 1$. Since positive label information is more informative than negative label information, as discussed in Section IV, we typically choose $\mu_1 > \mu_2$. By setting the partial derivative of $\mathcal{Q}(\mathbf{F})$ with respect to $\mathbf{F}$ to zero, we obtain the following optimal solution for $\mathbf{F}$:

$$\mathbf{F}^* = \left[\mathbf{L} + (\mu\mu_1 - \mu\mu_2)\,\mathbf{I}\right]^{-1} \left(\mu_1 \mathbf{Y}^+ - \mu\mu_2 \mathbf{Y}^-\right) \tag{18}$$

Finally, label assignment is performed according to (9).

Another possible straightforward approach for PNLPs could be treating the positive labels as negative ones, and then using the method introduced in Section IV. The disadvantage of this approach with respect to the proposed one is that, by combining the positive and the negative label information in the same label matrix Ψ, we assume that the significance of the positive and the negative label information is equivalent. However, as it was shown in paper, the significance of positive and negative labels is not equal. On the contrary, increased propagation accuracy is achieved when higher significance is given to the positive labels than to the negative ones.

Contrary to positive labels, negative label information appears more rarely in real world scenarios, e.g., in person identity LP on facial images extracted from movies. By knowing the movie from which each facial image was extracted and the actors that appear in the cast, we can prevent a facial image from being assigned the label of an actor that does not appear in the specific movie. In the other cases, negative label information can be imposed effectively on the data manually, according to the following procedure. First, LP is applied on the data by considering only positive label information, according to (4) and (5). As stated before, the values in $\mathbf{F}$ are an indication on the "certainty" with which the node is assigned a label. This means that nodes, in which the largest F_{ij} value is much larger than the second largest F_{ij} value, are more probably assigned the correct label, while nodes in which the two highest F_{ij} values are very close to each other, most probably lie in a "border" or "transition" region between two facial image classes. Label assignment to such nodes is more uncertain. The propagated labels to the nodes with the least certainty are examined, in order to form the set of negative labeled set. More specifically, for each node i, the difference between the two largest values in the ith row of $\mathbf{F}$ is computed. The q nodes with the smallest difference value are selected and their assigned label is examined. If the label is incorrect, then the node enters the negative labeled set, describing labeling constraints. Finally, the initial state matrix $\mathbf{Y}^-$ is updated with the negative labels and the LP is reperformed according to (17) and (5), this time considering both the positive and the negative label information. As it will be seen in the experiments, this choice for the negative labels increases significantly the classification accuracy of LP.

2.4 EXTENSION OF POSITIVE AND NEGATIVE LABEL PROPAGATIONS ON OUT OF SAMPLE DATA

The proposed PNLP frameworks, described by the regularization framework in (17), can be modified in order to assign labels to out of sample data (OSD) similar to [4]. The regularization framework (17) for the sample $\mathbf{x}_i \in \mathcal{X}$ is written as

$$\begin{aligned}\mathcal{Q}\left(\mathbf{f}_i\right) =& \frac{1}{2}\sum_{j=1}^{N} W_{ij}\left(\mathbf{f}_i - \mathbf{f}_j\right)^2 \\ &+ \frac{\mu}{2}\left[\mu_1\left(\mathbf{f}_i - \mathbf{y}_i^+\right)^2 - \mu_2\left(\mathbf{f}_i - \mathbf{y}_i^-\right)^2\right]\end{aligned} \tag{19}$$

where $\mathbf{f}_i, \mathbf{y}_i^+, \mathbf{y}_i^- \in \mathbb{R}^L$ is the ith row of matrix $\mathbf{F}$, $\mathbf{Y}^+$, and $\mathbf{Y}^- \in \mathbb{R}^{N\times L}$, respectively. When a new sample $\mathbf{X}$ is encountered for which negative label information $\mathbf{y}^- \in \mathbb{R}^L$ is available, the smoothness criterion becomes

$$\mathcal{Q}(\mathbf{f}(\mathbf{x})) = \frac{1}{2}\sum_{j=1}^{N} W\left(\mathbf{x}, \mathbf{x}_j\right)\left(\mathbf{f}(\mathbf{x}) - \mathbf{f}_j\right)^2 - \frac{\mu\mu_2}{2}\left(\mathbf{f}(\mathbf{x}) - \mathbf{y}^-\right)^2. \tag{20}$$

Since $\mathcal{Q}(\mathbf{f}(\mathbf{x}))$ is convex in $\mathbf{f}(\mathbf{x})$, it is minimized by setting $((\partial\mathcal{Q}(\mathbf{f}(\mathbf{x})))/(\partial\mathbf{f}(\mathbf{x}))) = 0$

$$\mathbf{f} = \frac{1}{\sum_{j=1}^{N} W\left(\mathbf{x}, \mathbf{x}_j\right) - \mu\mu_2}\left[\sum_{j=1}^{N} W\left(\mathbf{x}, \mathbf{x}_j\right)\mathbf{f}_j - \mu\mu_2\mathbf{y}^-\right] \tag{21}$$

We notice that the optimal score vector $\mathbf{f}$ is a linear combination of the score vectors of the training data and the negative label vector. Finally, label assignment is performed according to

$$l = \arg\max\{\mathbf{f}\} \tag{22}$$

3 Method Implementation details

3.1 Comparing with released source code

No related source codes are available. All codes were finished by myself.

3.2 Experimental environment setup

Hardware: Macbook pro(Apple M1 Pro)

Software: Mac OS, Matlab R2022a

3.3 Algorithm

The details of algorithm are as follows:

Algorithm 1 Positive and Negative Label Propagation

Input: Data $\mathbf{X} = [\mathbf{X}_L, \mathbf{X}_U]$, Positive Label $\mathbf{Y}^+$, Negative Label $\mathbf{Y}^-$, parameter $\mu > 0$, $0 < \mu_1 < 1$, $0 < \mu_2 < 1$, $\mu_1 + \mu_2 = 1$.

Output: optimal classification function $\mathbf{F}$

Calculate W by equation(1) and calculate $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$, where $\mathbf{D} = \mathrm{diag}_i\left\{\sum_{j=1:N} W_{ij}\right\}$

Calculate optimal $\mathbf{F}^*$ by equation(18)

label assignment is performed according to (12)

3.4 Coding

The Matlab code of PNLP is as follows:

```
function [F] = func_PNLP(XL, XU, YL, YU, sigma, miu, miu1, miu2)
    X = [XL, XU];
    n_total = size(X, 2);

    % Calculate XLX'
    W = zeros(n_total, n_total);
    k=10;
    knn_mat = func_getKNNMat(X,k);
    for i = 1:n_total
        for j = 1:k
            W(i, knn_mat(j, i)) = exp(-norm(X(:,i)-X(:,knn_mat(j, i)))^2/(sigma^2));
        end
    end
    D = diag(sum(W, 2));
    L = (D^(-1/2))*W*(D^(-1/2));

    % Calculate positive labels and negative labels
    Y_positive = [YL, YU];
    Y_negative = ~Y_positive;


    % Calculate the optimal F
    I = eye(size(L));
    F = miu*(L+miu*(miu1-miu2)*I)\(miu1*Y_positive-miu2*Y_negative)';
end
```

3.5 Experiment

For feature extraction and dimensional reduction, we used PCA to map original data to 300-D feature subspace firstly. For the selection of the parameters μ, μ_1 and μ_2, let $\mu = 10$, $\mu_1 = 0.3$ and $\mu_2 = 0.7$. In the ORL data set, we randomly select 5 samples from each class as training samples, and 2 of these 5 are labeled. In the AR data set, we randomly selected 10 samples from each class as training samples, and 4 of these 10 are labeled. In the Yale data set, we randomly select 6 samples from each class as training samples, and 2 of these 6 are labeled. In order to test the robustness of the algorithm, random occlusions with pixel sizes of 3 by 3 and 5 by 5 were added to all face data in the experiment respectively. Experiments on each dataset were repeated 10 times under the same size of occlusion. The results on different dataset are shown in table 1, table 2 and table 3.

Table 1: The results on ORL.

size of occlusion	means	std	max	min
0×0	82.00%	3.97%	85.00%	78.00%
3×3	78.30%	3.68%	85.50%	75.50%
5×5	79.25%	3.34%	85.50%	74.50%

Table 2: The results on AR.

size of occlusion	means	std	max	min
0×0	68.38%	2.94%	70.83%	66.17%
3×3	68.94%	2.56%	71.75%	66.75%
5×5	69.25%	0.55%	69.75%	68.67%

Table 3: The results on Yale.

size of occlusion	means	std	max	min
0×0	81.87%	4.37%	89.33%	76.00%
3×3	76.73%	4.94%	88.00%	73.33%
5×5	79.40%	6.80%	88.00%	65.33%

From the results in the above table, we can see that even if random occlusion is added, the recognition rate of the algorithm does not drop significantly, indicating that the algorithm has strong robustness.

4 Conclusion

This report introduces a semi-supervised learning algorithm —- Positive and Negative label propagation. A novel method has been presented that introduces the problem of negative LP in the task of single-graph LP. More specifically, the state-of-the-art LP methods propagate information of the form "the sample i should be assigned the label k." The proposed method extends the state-of-the-art framework by considering additional information of the form "the sample i should not be assigned the label k." Moreover, a method for selecting the negative labels in cases when they are not inherent from the data structure has been introduced. The effectiveness of the proposed framework becomes more significant when the data contain inherent negative label information.

References

[1] X. J. Zhu, "Semi-supervised learning literature survey," 2005.

[2] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," vol. 16, 2003.

[3] X. He and P. Niyogi, "Locality preserving projections," *Advances in neural information processing systems*, vol. 16, 2003.

[4] F. Wang and C. Zhang, "Label propagation through linear neighborhoods," in *Proceedings of the 23rd international conference on Machine learning*, pp. 985–992, 2006.