SEMI-SUPERVISED CLASSIFICATION VIA BOTH LABEL AND SIDE INFORMATION

Rui Zhang, Feiping Nie*, and Xuelong Li

School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL)
Northwestern Polytechnical University, Xi’an 710072, Shaanxi, China.

ABSTRACT
As for the semi-supervised learning, both label and side information serve as pretty significant indicators for the classification. However, majority of the associated works only focus on one side of the road. In other words, either the label information or the side information is utilized instead of taking both of them into consideration simultaneously. To address the referred defect, we propose a graph-based semi-supervised learning (GSL) problem via building the intrinsic graph and the penalty graph upon both label and side information. To efficiently unravel the proposed GSL problem, a novel quadratic trace ratio (QTR) method is proposed based on solving the associated QTR problem, which is the equivalent counterpart of the GSL problem. Besides, a parameter-free similarity is further derived and utilized. Consequently, a novel semi-supervised classification (SC) algorithm can be summarized by virtue of the proposed GSL problem and QTR method.

Index Terms— soft label, side information, graph-based semi-supervised learning, quadratic trace ratio problem.

1. INTRODUCTION
Recently, graph-based semi-supervised learning[1, 2, 3] has aroused strong motivations of multiple researches in machine learning and pattern recognition. Graph-based methods [4] usually initiate a graph path, where the data points are divided into labeled and unlabeled categories with the edges reflecting the pairwise similarity. Under the assumption that the connected points tend to belong to the same class, the labels could effectively propagate via the proposed graph. Besides, graph-based methods[5, 6, 7] could always benefit from the superior statistical properties, which are closely related to the spectral graph theory.

There are numerous pivotal works concerning the graph-based semi-supervised learning. In [8], Zhu et al. proposed the label propagation (LP) method based on investigating the graph-based semi-supervised learning problem in terms of the harmonic Gaussian random field model. In [9], Zhou et al. proposed the learning with local and global consistency (LLGC) method based on solving the semi-supervised learning problem under the smooth structure. Admittedly, all these works impart the meaningful contributions towards certain optimization problems. However, most of the semi-supervised methods utilize either the label information or the side information instead of exploiting both at the same time.

To address the referred defect, an original semi-supervised classification (SC) method is proposed by utilizing both label and side information. We contribute to this paper in the following aspects. 1. Via the intrinsic graph and the penalty graph upon both label and side information, the graph-based semi-supervised learning (GSL) problem is proposed as a bi-objective graph optimization. 2. To solve the proposed GSL problem, a novel quadratic trace ratio (QTR) method is derived by introducing a characteristic function. 3. The SC method can be summarized by virtue of the proposed GSL problem and QTR method with better classification results on both synthetic and real databases. 4. A novel parameter-free similarity is further derived and utilized.

2. PARAMETER-FREE SIMILARITY
Suppose input data \( X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d \times n} \) with feature dimension \( d \) and data number \( n \), then the symmetric adjacent matrix \( A = [a_{ij}]_{n \times n} \) and the diagonal degree matrix \( D = \mathbb{1}_{n \times n} (d_{ii} = \sum_{j=1}^{n} a_{ij}) \) can be constructed. To achieve the parameter-free similarity, we introduce the following optimization w.r.t. \( a_{i} \) as

\[
\min_{a_{i}1=1.0 \leq a_{i} \leq 1} \quad \frac{\min}{\text{Tr}(X(D-A)X^T)} + \sum_{i,j=1}^{n} \left( \frac{\gamma_{i}}{a_{ij}^2} \right) \quad (1)
\]

where \( a_{i} \) is the \( i \)-th column of the adjacent matrix \( A \) and \( \gamma_{i} \) is the regularization parameter with \( 1 = [1, 1, \ldots, 1]^T \in \mathbb{R}^{n \times 1} \). Apparently, the problem (1) could be reformulated into

\[
\min_{a_{i}1=1.0 \leq a_{i} \leq 1} \quad \frac{1}{2} \sum_{i,j=1}^{n} \left( a_{ij} \| x_{i} - x_{j} \|_2^2 + \gamma_{i} a_{ij}^2 \right)
\]

\[
\Rightarrow \min_{a_{i}1=1.0 \leq a_{i} \leq 1} \quad \frac{1}{2} \sum_{i=1}^{n} \left( a_{i} + \frac{e_{i} \| x_{i} - x_{j} \|_2^2}{2\gamma_{i}} \right) \quad (2)
\]

where \( e_{i} \in \mathbb{R}^{n \times 1} \) serves as a column vector with its \( j \)-th element being \( e_{ij} = \| x_{i} - x_{j} \|_2^2 \). Note that the problem (2) is
independent between each two terms, thus we can solve each i-th term of the problem (2) individually as
\[
\min_{a_i^T 1=1, 0 \leq a_i \leq 1} \frac{1}{2} ||a_i + \frac{e_i}{2\gamma_i}||_2^2. \tag{3}
\]

Therefore, Lagrangian function of the problem (3) could be represented as
\[
\mathcal{L}(a_i, \eta, \beta_j) = \frac{1}{2} ||a_i + \frac{e_i}{2\gamma_i}||_2^2 - \eta(a_i^T 1 - 1) - \beta_i^T a_i
\]
where \( \eta \) and \( \beta_i \geq 0 \) are Lagrangian multipliers.

Accordingly, the KKT condition could be illustrated as
\[
a_{ij} = \left( -\frac{e_{ij}}{2\gamma_i} + \eta \right)_+ \tag{4}
\]

For practical purpose, we target at obtaining a sparse similarity matrix \( A \). In other words, only \( k \) nearest neighbors of each data point are taken into consideration. Without loss of generality, we could assume \( e_{ij} \leq e_{i2} \cdots \leq e_{in}, \forall i \). If vector \( a_i \) in (4) has exact \( k \) nonzero numbers, we could infer that
\[
\begin{align*}
\frac{1}{2} a_i^T 1 - 1 &< 0 \\
\frac{1}{2} a_i^T 1 - 1 &\leq \eta \Rightarrow \eta \geq \frac{1}{2} a_i^T 1 - 1.
\end{align*}
\]

Due to the constraint \( a_i^T 1 = 1 \) and Eq. (4), we have
\[
\sum_{j=1}^{k} \left( -\frac{e_{ij}}{2\gamma_i} + \eta \right) = 1 \Rightarrow \eta = \frac{1}{k} + \frac{1}{2k\gamma_i} \sum_{j=1}^{k} e_{ij}. \tag{5}
\]

Based on the results in (5) and (6), the following inequality of \( \gamma_i \) could be derived as
\[
\frac{k}{2} e_{ik} - \frac{1}{2} \sum_{j=1}^{k} e_{ij} < \gamma_i \leq \frac{k}{2} e_{i(k+1)} - \frac{1}{2} \sum_{j=1}^{k} e_{ij}. \tag{7}
\]

By virtue of Eq. (7), we could set \( \gamma_i = \frac{k}{2} e_{i(k+1)} - \frac{1}{2} \sum_{j=1}^{k} e_{ij} \) such that an optimal solution \( a_i \) is achieved with exact \( k \) nonzero values. Accordingly, the must-link similarity matrix \( A^w \) and the cannot-link similarity matrix \( A^b \) can be specifically constructed as
\[
a^w_{ij} = \begin{cases} \infty & (x_i, x_j) \text{ in same class} \\ (-\frac{e_{ij}}{2\gamma_i} + \eta)_+ & \text{otherwise} \end{cases}
\]
and
\[
a^b_{ij} = \begin{cases} 1 & (x_i, x_j) \text{ in different class} \\ 0 & \text{otherwise} \end{cases}
\]

where \( a^w_{ij} \) and \( a^b_{ij} \) are the \( ij \)-th elements of \( A^w \) and \( A^b \), respectively. Besides, \( \infty \) in (8) stands for a very large number such that the related labels are forced to be the same when the data \( x_i \) and \( x_j \) belong to the same class.

\section{3. GRAPH OPTIMIZATION CONCERNING SOFT LABEL AND SIDE INFORMATION}

Suppose that data \( X \in \mathbb{R}^{d \times n} \) are distributed into \( c \) different classes, then we try to utilize both label and side information for better classification. By virtue of the corresponding soft label matrix \( Y \in \mathbb{R}^{n \times c} \) and the side information concerning the pairwise constraints, both the intrinsic graph \( G = \{Y^T, A^w\} \) and the penalty graph \( G^p = \{Y^T, A^b\} \) can be constructed. Moreover, the soft label \( Y = [y_1, y_2, \ldots, y_n]^T \) retains the probability of the possible case \( x_i \in j \)-th class, \( \forall j \) in the related soft label \( y_i \in \mathbb{R}^{c \times 1}, (i = 1, 2, \ldots, n) \) such that the label matrix \( Y \) is free from the traditional binary constraint.

Generally speaking, the classification problem is to minimize the intrinsic graph problem \( G \) with maximizing the penalty graph problem \( G^p \) simultaneously. Therefore, the classification problem can be further represented as the following bi-objective graph optimization
\[
\begin{align*}
\min_{Y} & \sum_{i,j} a^w_{ij} ||y_i - y_j||_2^2 = \min_{Y} 2Tr(Y^TL^wY) \\
\max_{Y} & \sum_{i,j} a^b_{ij} ||y_i - y_j||_2^2 = \max_{Y} 2Tr(Y^TL^bY)
\end{align*}
\tag{9}
\]

where must-link graph Laplacian \( L^w = D^w - A^w \in \mathbb{R}^{n \times n} \) and cannot-link graph Laplacian \( L^b = D^b - A^b \in \mathbb{R}^{n \times n} \) with \( D^w = \text{diag}(\sum_{j=1}^{n} a^w_{1j}, \sum_{j=1}^{n} a^w_{2j}, \ldots, \sum_{j=1}^{n} a^w_{nj}) \) and \( D^b = \text{diag}(\sum_{j=1}^{n} a^b_{1j}, \sum_{j=1}^{n} a^b_{2j}, \ldots, \sum_{j=1}^{n} a^b_{nj}) \).

Accordingly, the problem (9) can be reformulated into
\[
\begin{align*}
\min_{Y} & \frac{Tr([y_1, y_2, \ldots, y_n]^T L^w[y_1, y_2, \ldots, y_n]^T)}{Tr([y_1, y_2, \ldots, y_n]^T L^b[y_1, y_2, \ldots, y_n]^T)}
\end{align*}
\tag{10}
\]

\section{4. GRAPH-BASED SEMI-SUPERVISED LEARNING}

The semi-supervised learning implies that part of the labels for the data \( X \) have already been identified i.e. the soft label matrix \( Y = [Y_l; F_u] \in \mathbb{R}^{n \times c} \) with labeled matrix \( Y_l \in \mathbb{R}^{n_l \times c} \) and unlabeled matrix \( F_u \in \mathbb{R}^{n_u \times c} \) satisfying \( n_l + n_u = n \).

Specifically, the labeled matrix \( Y_l \) is binary since each labeled data belongs to only one class with 100% probability.

Based on the problem (9) and (10), the graph-based semi-supervised learning (GSL) problem can be represented as
\[
\begin{align*}
\min_{F_u} & \frac{Tr([Y_l; F_u]^T L^w [Y_l; F_u])}{Tr([Y_l; F_u]^T L^b [Y_l; F_u])} \\
= & \min_{F_u} \frac{Tr([Y_l; F_u]^T [L^w_{ll} L^w_{lu} L^w_{ul} L^w_{uu}] [Y_l; F_u])}{Tr([Y_l; F_u]^T [L^b_{ll} L^b_{lu} L^b_{ul} L^b_{uu}] [Y_l; F_u])}
\end{align*}
\tag{11}
\]

where \( [L^w_{ll} \in \mathbb{R}^{n_l \times n_l}, L^w_{lu} \in \mathbb{R}^{n_l \times n_u}, L^w_{ul} \in \mathbb{R}^{n_u \times n_l}, L^w_{uu} \in \mathbb{R}^{n_u \times n_u}] \) and \( [L^b_{ll} \in \mathbb{R}^{n_l \times n_l}, L^b_{lu} \in \mathbb{R}^{n_l \times n_u}, L^b_{ul} \in \mathbb{R}^{n_u \times n_l}, L^b_{uu} \in \mathbb{R}^{n_u \times n_u}] \).
Input: $A, B, C, D, e$ and $f$ defined in (12).

Output: $Q$.

1. Initialize $p = 1$, $\lambda_1 = 0$ and $\lambda_2$ such that $A - \lambda_2 B$ is positive definite;
2. while $p > 0$ do
   3. Update $\lambda \leftarrow \frac{\lambda_1 + \lambda_2}{2}$;
   4. Update $Q \leftarrow (A - \lambda B)^{-1}(\lambda D - C)$;
   5. Update $p \leftarrow Tr(Q^T (A - \lambda B)Q) + 2Tr(Q^T (C - \lambda D)) + (e - \lambda f)$;
   6. if $p > 0$ then
      7. Replace $\lambda_1 \leftarrow \lambda$;
   8. end
9. end
10. while not converge do
    11. Update $Q \leftarrow (A - \lambda B)^{-1}(\lambda D - C)$;
    12. Update $\lambda \leftarrow \frac{Tr(Q^T AQ) + 2Tr(Q^T C) + e}{Tr(Q^T BQ) + 2Tr(Q^T D) + f}$;
13. end
14. return $Q$;

Algorithm 1: Quadratic trace ratio (QTR) method

$L_{uu}^{l} \in \mathbb{R}_{n_u \times n_u}$ are the block matrix representations for
the must-link graph Laplacian $L_{u}$ and the cannot-link graph
Laplacian $L_{l}^{u}$, respectively.

Apparently, the GSL problem (11) is equivalent to the fol-
lowing quadratic trace ratio (QTR) problem
\[
\min_{Q \in \mathbb{R}_{n_u \times n_u}} Tr(Q^T AQ) + 2Tr(Q^T C) + e
\]
\[
\lambda Tr(Q^T BQ) + 2Tr(Q^T D) + f
\]
where $A = L_{uu}^{w}$, $B = L_{uu}^{l}$, $C = L_{ul}^{w}Y_{l}$, $D = L_{ul}^{l}Y_{l}$, $e =
Tr(Y_{l}^T L_{ul}^{w}Y_{l})$ and $f = Tr(Y_{l}^T L_{ul}^{l}Y_{l})$ with $Tr([Y_{l}; Q]^T L_{ul}^{w}L_{ul}^{l}; Q]) > 0$.

To solve the QTR problem (12), we introduce the character-
istic function $p(\lambda)$ as
\[
p(\lambda) = \min_{Q} (Tr(Q^T AQ) + 2Tr(Q^T C) + e) -
\lambda (Tr(Q^T BQ) + 2Tr(Q^T D) + f)
\]
where $\lambda \leftarrow \frac{Tr(Q^T AQ) + 2Tr(Q^T C) + e}{Tr(Q^T BQ) + 2Tr(Q^T D) + f}$ is to be updated in the
algorithm 1.

Accordingly, we could infer that
\[
p(\lambda) \Rightarrow \min_{Q} Tr(Q^T (A - \lambda B)Q) + 2Tr(Q^T (C - \lambda D))
\]
\[
\Rightarrow Q = (A - \lambda B)^{-1}(\lambda D - C).
\]

(14)

Based on $p(\lambda)$ in (13) and result in (14), the quadratic trace ratio (QTR) method can be outlined in the algorithm 1. Besides, the algorithm 1 monotonically converges to the global opti-

umum of the QTR problem (12) with quadratic convergence rate in [10].

Based on the proposed GSL problem and QTR method, the semi-supervised classification (SC) method can be sum-
marized in the algorithm 2.

5. EXPERIMENTAL RESULTS

We divide the experiment into two parts concerning the syn-
thetic database and the real database to show the effectiveness of our method. All the comparisons are performed under the same priori knowledge.

5.1. Synthetic database

We first utilize two-spirals synthetic database to compare the classification results among the LP [8] method, the LLGC [9] method and the proposed SC method in the figure 1. We fur-
ther compare the classification results of the methods men-

![Two-spirals synthetic data](image1)

(a) Raw data with priori knowledge
(b) LP method
(c) LLGC method
(d) SC method(our)

Fig. 1. The classification comparison is performed for the
LP[8] method, the LLGC[9] method and the proposed SC
method under the two-spirals synthetic data.

![Three-rings synthetic data](image2)

(a) LP method
(b) LLGC method
(c) SC method(our)

Fig. 2. The classification comparison is performed for the
LP[8] method, the LLGC[9] method and the proposed SC
method under the three-rings synthetic data.
Table 1. The comparison of the classification accuracy under different labeled data.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>USPS</th>
<th>40 labeled data</th>
<th>60 labeled data</th>
<th>80 labeled data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Acc. (%)</td>
<td>Dev. (%)</td>
<td>Acc. (%)</td>
<td>Dev. (%)</td>
</tr>
<tr>
<td>k-NN[11]</td>
<td>64.34 ± 0.82</td>
<td>66.59 ± 0.95</td>
<td>73.87 ± 0.79</td>
<td></td>
</tr>
<tr>
<td>SVM[12]</td>
<td>62.99 ± 0.63</td>
<td>68.28 ± 1.41</td>
<td>74.10 ± 0.91</td>
<td></td>
</tr>
<tr>
<td>LP[8]</td>
<td>77.50 ± 0.66</td>
<td>79.58 ± 0.34</td>
<td>81.72 ± 0.52</td>
<td></td>
</tr>
<tr>
<td>LLGC[9]</td>
<td>81.92 ± 1.97</td>
<td>83.33 ± 2.06</td>
<td>84.50 ± 1.68</td>
<td></td>
</tr>
<tr>
<td>SC(our)</td>
<td><strong>83.35 ± 0.68</strong></td>
<td><strong>85.87 ± 0.54</strong></td>
<td><strong>87.48 ± 0.99</strong></td>
<td></td>
</tr>
</tbody>
</table>

Input: input data $X$ and labeled matrix $Y_i$.

Output: the binary label matrix $Y$.

1. Initialize $A^w$ and $A^b$ defined in (8) via the given pair constraints $(x_i, x_j)$;
2. Calculate $F_u$ via the QTR method in the algorithm 1;
3. $Y_u \leftarrow F_u$, where $Y_u$ is the binary label under the soft label $F_u$ (i.e. $j^* = \arg \max_{1 \leq j \leq c} F_u(i,j)$, $\forall i$ such that $Y_u(i,j^*) = 1$ with $\sum_j Y_u(i,j) = 1$);
4. return $Y = [Y_i; Y_u]$;

Algorithm 2: Semi-supervised classification (SC) method under the proposed GSL problem and QTR method

---

**Fig. 3.** The error rate comparisons are performed for $k$-NN[11], SVM[12], LP[8], LLGC[9] and SC(our) via different percentages taken by the labeled data under 6 recognition datasets.

---

**6. CONCLUDING REMARKS**

We propose a semi-supervised classification method by simultaneously utilizing both the label and the side information. Besides, the related semi-supervised learning problem is represented as a bi-objective graph optimization. To solve the associated semi-supervised learning problem, a novel quadratic trace ratio method is derived by introducing the characteristic function. Eventually, we perform extensive experiments on both synthetic and real datasets, which illustrate that the proposed semi-supervised classification method is superior than the conventional methods.
7. REFERENCES


