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 $D \in \mathbb{R}^{n\times n}$, $(d_{ii} = \sum_{j=1}^{n} a_{ij})$ can be constructed. To achieve the parameter-free similarity, we introduce the following optimization w.r.t. a_i as

$$\min_{a_i^T\mathbf{1}=1, 0\le a_i\le 1} Tr(X(D-A)X^T) + \sum_{i,j=1}^{n} (\frac{\gamma_i}{2} a_{ij}^2) \qquad (1)$$

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

$$\begin{aligned} &\min_{a_i^T\mathbf{1}=1, 0\le a_i\le 1} \frac{1}{2}\sum_{i,j=1}^{n} (a_{ij}\|x_i - x_j\|_2^2 + \gamma_i a_{ij}^2) \\ \Rightarrow &\min_{a_i^T\mathbf{1}=1, 0\le a_i\le 1} \sum_{i=1}^{n} \frac{1}{2}\|a_i + \frac{e_i}{2\gamma_i}\|_2^2 \end{aligned} \qquad (2)$$

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

*Corresponding author. Email: feipingnie@gmail.com.

independent between each two terms, thus we can solve each i-th term of the problem (2) individually as

$$\min_{a_i^T\mathbf{1}=1,0\le a_i\le 1} \frac{1}{2}\|a_i+\frac{e_i}{2\gamma_i}\|_2^2. \tag{3}$$

Therefore, Lagrangian function of the problem (3) could be represented as

$$\mathscr{L}(a_i,\eta,\beta_i)=\frac{1}{2}\|a_i+\frac{e_i}{2\gamma_i}\|_2^2-\eta(a_i^T\mathbf{1}-1)-\beta_i^T a_i$$

where η and $\beta_i\ge 0$ are Lagrangian multipliers.

Accordingly, the KKT condition could be illustrated as

$$a_{ij}=(-\frac{e_{ij}}{2\gamma_i}+\eta)_+. \tag{4}$$

For practical purpose, we target at obtaining a sparse similarity A. In other words, only k nearest neighbors of each data point are taken into consideration. Without loss of generality, we could assume $e_{i1}\le e_{i2}\cdots\le e_{in},\forall i$. If vector a_i in (4) has exact k nonzero numbers, we could infer that

$$\begin{cases} a_{ik}>0\Rightarrow -\frac{e_{ik}}{2\gamma_i}+\eta>0 \\ a_{i(k+1)}\le 0\Rightarrow -\frac{e_{i(k+1)}}{2\gamma_i}+\eta\le 0 \end{cases}. \tag{5}$$

Due to the constraint $a_i^T\mathbf{1}=1$ and Eq. (4), we have

$$\sum_{j=1}^{k}(-\frac{e_{ij}}{2\gamma_i}+\eta)=1\Rightarrow\eta=\frac{1}{k}+\frac{1}{2k\gamma_i}\sum_{j=1}^{k}e_{ij}. \tag{6}$$

Based on the results in (5) and (6), the following inequality of γ_i could be derived as

$$\frac{k}{2}e_{ik}-\frac{1}{2}\sum_{j=1}^{k}e_{ij}<\gamma_i\le\frac{k}{2}e_{i(k+1)}-\frac{1}{2}\sum_{j=1}^{k}e_{ij}. \tag{7}$$

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

$$a_{ij}^w=\begin{cases}\infty & (x_i,x_j)\ in\ same\ class\\ (-\frac{e_{ij}}{2\gamma_i}+\eta)_+ & otherwise\end{cases}$$

and

$$a_{ij}^b=\begin{cases}1 & (x_i,x_j)\ in\ different\ class\\ 0 & otherwise\end{cases} \tag{8}$$

where a_{ij}^w and a_{ij}^b are the ij-th elements of A^w and A^b, respectively. Besides, ∞ 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.

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

$$\begin{cases}\min\limits_{Y}\sum\limits_{i,j}a_{ij}^w\|y_i-y_j\|_2^2=\min\limits_{Y}2Tr(Y^TL^wY)\\ \max\limits_{Y}\sum\limits_{i,j}a_{ij}^b\|y_i-y_j\|_2^2=\max\limits_{Y}2Tr(Y^TL^bY)\end{cases} \tag{9}$$

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 $D^w=diag(\sum_{j=1}^n a_{1j}^w,\sum_{j=1}^n a_{2j}^w,\ldots,\sum_{j=1}^n a_{nj}^w)$ and $D^b=diag(\sum_{j=1}^n a_{1j}^b,\sum_{j=1}^n a_{2j}^b,\ldots,\sum_{j=1}^n a_{nj}^b)$.

Accordingly, the problem (9) can be reformulated into

$$\min_{Y}\frac{Tr([y_1,y_2,\ldots,y_n]L^w[y_1,y_2,\ldots,y_n]^T)}{Tr([y_1,y_2,\ldots,y_n]L^b[y_1,y_2,\ldots,y_n]^T)}. \tag{10}$$

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

$$\begin{aligned}&\min_{F_u}\frac{Tr([Y_l;F_u]^TL^w[Y_l;F_u])}{Tr([Y_l;F_u]^TL^b[Y_l;F_u])}\\ &=\min_{F_u}\frac{Tr(\begin{bmatrix}Y_l\\F_u\end{bmatrix}^T\begin{bmatrix}L_{ll}^w & L_{lu}^w\\ L_{ul}^w & L_{uu}^w\end{bmatrix}\begin{bmatrix}Y_l\\F_u\end{bmatrix})}{Tr(\begin{bmatrix}Y_l\\F_u\end{bmatrix}^T\begin{bmatrix}L_{ll}^b & L_{lu}^b\\ L_{ul}^b & L_{uu}^b\end{bmatrix}\begin{bmatrix}Y_l\\F_u\end{bmatrix})}\end{aligned} \tag{11}$$

where $[L_{ll}^w\in\mathbb{R}^{n_l\times n_l},L_{lu}^w\in\mathbb{R}^{n_l\times n_u};L_{ul}^w\in\mathbb{R}^{n_u\times n_l},L_{uu}^w\in\mathbb{R}^{n_u\times n_u}]$ and $[L_{ll}^b\in\mathbb{R}^{n_l\times n_l},L_{lu}^b\in\mathbb{R}^{n_l\times n_u};L_{ul}^b\in\mathbb{R}^{n_u\times n_l},$

Input: A, B, C, D, e and f defined in (12).
Output: Q.

Initialize $p = 1$, $\lambda_1 = 0$ and λ_2 such that $A - \lambda_2 B$ is positive definite;
while $p > 0$ **do**
  Update $\lambda \leftarrow \frac{\lambda_1+\lambda_2}{2}$;
  Update $Q \leftarrow (A - \lambda B)^{-1}(\lambda D - C)$;
  Update $p \leftarrow Tr(Q^T(A - \lambda B)Q) + 2Tr(Q^T(C - \lambda D)) + (e - \lambda f)$;
  if $p > 0$ **then**
    Replace $\lambda_1 \leftarrow \lambda$;
  end
end
while *not converge* **do**
  Update $Q \leftarrow (A - \lambda B)^{-1}(\lambda D - C)$;
  Update $\lambda \leftarrow \frac{Tr(Q^T AQ)+2Tr(Q^T C)+e}{Tr(Q^T BQ)+2Tr(Q^T D)+f}$;
end
return Q;

Algorithm 1: Quadratic trace ratio (QTR) method

$L^b_{uu} \in \mathbb{R}^{n_u \times n_u}]$ are the block matrix representations for the must-link graph Laplacian L^w and the cannot-link graph Laplacian L^b, respectively.

Apparently, the GSL problem (11) is equivalent to the following quadratic trace ratio (QTR) problem

$$\min_{Q \in \mathbb{R}^{n_u \times c}} \frac{Tr(Q^T AQ) + 2Tr(Q^T C) + e}{Tr(Q^T BQ) + 2Tr(Q^T D) + f} \tag{12}$$

where $A = L^w_{uu}$, $B = L^b_{uu}$, $C = L^w_{ul}Y_l$, $D = L^b_{ul}Y_l$, $e = Tr(Y_l^T L^w_{ll} Y_l)$ and $f = Tr(Y_l^T L^b_{ll} Y_l)$ with $Tr([Y_l; Q]^T L^b [Y_l; Q]) > 0$.

To solve the QTR problem (12), we introduce the characteristic function $p(\lambda)$ as

$$\begin{aligned} p(\lambda) = \min_Q (Tr(Q^T AQ) + 2Tr(Q^T C) + e) - \\ \lambda(Tr(Q^T BQ) + 2Tr(Q^T D) + f) \end{aligned} \tag{13}$$

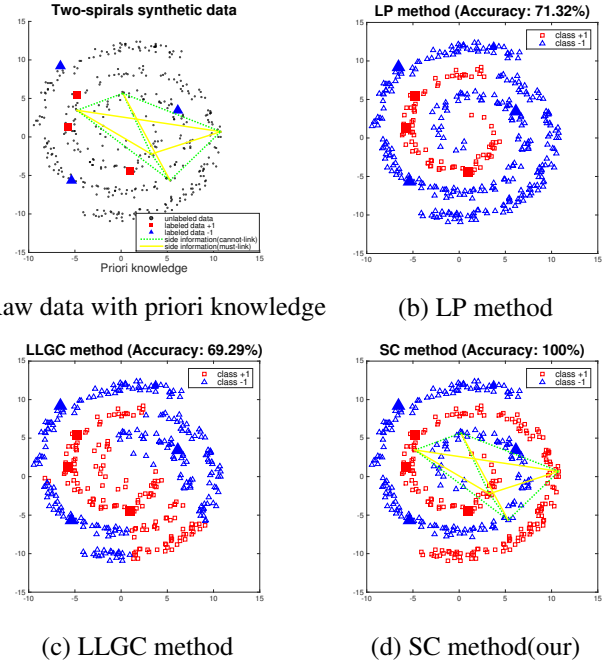
where $\lambda \leftarrow \frac{Tr(Q^T AQ)+2Tr(Q^T C)+e}{Tr(Q^T BQ)+2Tr(Q^T D)+f}$ is to be updated in the algorithm 1.

Accordingly, we could infer that

$$\begin{aligned} p(\lambda) \Rightarrow \min_Q Tr(Q^T(A - \lambda B)Q) + 2Tr(Q^T(C - \lambda D)) \\ \Rightarrow Q = (A - \lambda B)^{-1}(\lambda D - C). \end{aligned} \tag{14}$$

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 optimum 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 summarized in the algorithm 2.



(a) Raw data with priori knowledge (b) LP method

(c) LLGC method (d) SC method(our)

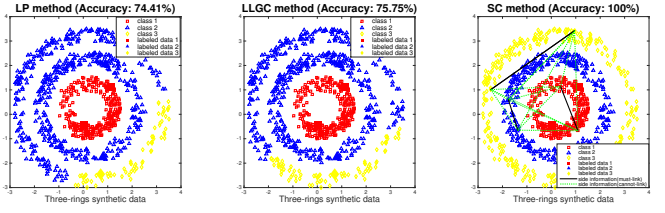
Fig. 1. The classification comparison is performed for the LP[8] method, the LLGC[9] method and the proposed SC method under the two-spirals synthetic data.

5. EXPERIMENTAL RESULTS

We divide the experiment into two parts concerning the synthetic 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 further compare the classification results of the methods men-



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

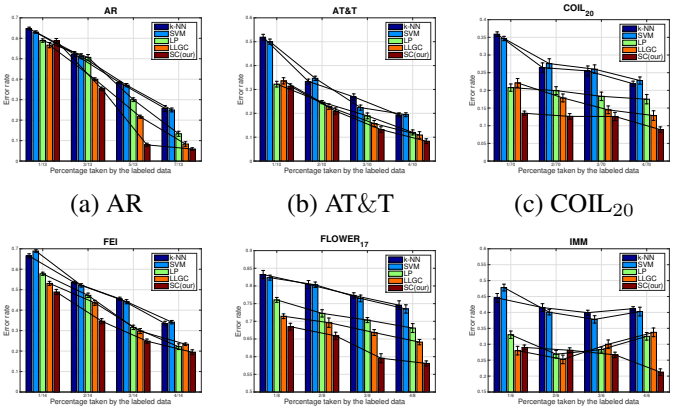
Fig. 2. The classification comparison is performed for the LP[8] method, the LLGC[9] method and the proposed SC method under the three-rings synthetic data.

Table 1. The comparison of the classification accuracy under different labeled data.

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

Input: input data X and labeled matrix Y_l.
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. $j^* = \arg\max_{1\le j\le c} F_u(i,j), \forall i$ such that $Y_u(i,j^*) = 1$ with $\sum_j Y_u(i,j) = 1$);
4 return $Y = [Y_l; Y_u]$;

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



(a) AR (b) AT&T (c) COIL_{20}

(d) FEI (e) FLOWER_{17} (f) IMM

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.

tioned above on the three-rings synthetic database in the figure 2. As for the LP [8] method and the LLGC [9] method, the optimal classification results are recorded via tuning the parameter. Accordingly, the classification results are illustrated in the figure 1 and 2. Therefore, we could conclude that:

1. From the figure 1 and 2, we could observe that the proposed SC method could achieve the optimal classification results based on utilizing both label and side information.

2. From the figure 1 and 2, we notice that the SC method performs better than the LP [8] method and the LLGC [9] method.

5.2. Real database

We use 7 benchmark recognition datasets to compare the classification accuracy under the k-NN [11] method, the SVM [12] method, the LP [8] method, the LLGC [9] method and the proposed SC method. As for the k-NN [11] method and the SVM [12] method, the classification results can be obtained by treating the labeled data as the training set and the unlabeled data as the test set. In the figure 3, we choose 6 datasets as AR, AT&T, COIL_{20}, FEI, FLOWER_{17} and IMM for the classification comparison with equal labeled data shared by each class. In the table 1, we randomly choose the labeled data distributed in the datasets USPS. Accordingly, the following conclusions can be drawn.

1. In the figure 3 and table 1, we could observe that the proposed SC method performs much better than other approaches on the classification accuracy with minor exceptions.

2. From the table 1, the LLGC [9] method pursues the classification accuracy with larger deviation due to its dependence on the regularization parameter.

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

[1] J. Yu, M. Wang, and D. Tao, "Semi-supervised multiview distance metric learning for cartoon synthesis," *IEEE Transactions on Image Processing*, vol. 21, no. 11, pp. 4636–4648, 2012.

[2] X. Cai, F. Nie, W. Cai, and H. Huang, "Heterogeneous image features integration via multi-modal semi-supervised learning model," in *IEEE*. International Conference on Computer Vision, 2013, pp. 1737–1744.

[3] F. Nie, X. Wang, M.I. Jordan, and H. Huang, "The constrained laplacian rank algorithm for graph-based clustering," in *Proceedings*. The 30th AAAI Conference on Artificial Intelligence, 2016.

[4] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, 2007.

[5] X. Cai, F. Nie, H. Huang, and F. Kamangar, "Heterogeneous image feature integration via multi-modal spectral clustering," in *IEEE*. Conference on Computer Vision and Pattern Recognition, 2011, vol. 4, pp. 1977–1984.

[6] T. Xia, D. Tao, T. Mei, and Y. Zhang, "Multi-view spectral embedding," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, no. 6, pp. 1438–1446, 2010.

[7] F. Nie, J. Li, and X. Li, "Parameter-free auto-weighted multiple graph learning: A framework for multi-view clustering and semi-supervised classification," in *Proceedings*. The 25th International Joint Conference on Artificial Intelligence, 2016.

[8] D. Zhu, Z. Ghahramani, and J.D. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *Proceedings*. The 20th International Conference on Machine Learning, 2003, pp. 912–919.

[9] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proceedings*. Advances in Neural Information Processing Systems, 2004, vol. 16, pp. 321–328.

[10] Y. Jia, F. Nie, and C. Zhang, "Trace ratio problem revisited," *IEEE Transactions on Neural Networks*, vol. 20, no. 4, pp. 729–735, 2009.

[11] O. Kramer, *Dimensionality Reduction with Unsupervised Nearest Neighbors*, Springer-Verlag, Berlin, Heidelberg, Germany, 2013.

[12] M.A. Hearst, "Semi-supervised mult-iview distance metric learning for cartoon synthesis," *IEEE Intelligent Systems*, vol. 13, no. 4, pp. 18–28, 1998.