Learning Strictly Orthogonal p-Order Nonnegative Laplacian Embedding via Smoothed Iterative Reweighted Method*

*The supplementary of this paper is supplied at https://www.dropbox.com/s/2rcbcyh5ikpsi3u/IJCAI__19_appendix.pdf?dl=0

Appendix: Learning Strictly Orthogonal p-Order Nonnegative Laplacian Embedding via Smoothed Iterative Reweighted Method

1 Proof of Theorem and Lemmas

Theorem 1 *The Algorithm* **??** *will monotonically decrease the objective of the problem (***??***) in each iteration until the algorithm converges.*

Before proving the convergence of the Algorithm **??**, we first introduce several lemmas:

Lemma 1 *For any $\sigma > 0$, the following inequality holds when $0 < p \leq 2$:*

$$\frac{p}{2}\sigma - \sigma^{\frac{p}{2}} + \frac{2-p}{2} \geq 0 \ . \tag{1}$$

Proof. Denote $f(\sigma) = p\sigma - 2\sigma^{\frac{p}{2}} + 2 - p$, we have the following derivatives:

$$f'(\sigma) = p(1 - \sigma^{\frac{p-2}{2}}), \quad \text{and} \quad f''(\sigma) = \frac{p(2-p)}{2}\sigma_i^{\frac{p-4}{2}}.$$

Obviously, when $\sigma > 0$ and $0 < p \leq 2$, then $f''(\sigma) \geq 0$ and $\sigma = 1$ is the only point that $f'(\sigma) = 0$. Note that $f(1) = 0$, thus when $\sigma > 0$ and $0 < p \leq 2$, then $f(\sigma) \geq 0$, which indicates Eq.(1). □

Lemma 2 *[?] For any positive definite matrices $\tilde{M}, M$ with the same size, suppose the eigen-decomposition $\tilde{M} = U\Sigma U^T$, $M = V\Lambda V^T$, where the eigenvalues in Σ is in increasing order and the eigenvalues in Λ is in decreasing order. Then the following inequality holds when $0 < p \leq 2$:*

$$Tr(\tilde{M}M) \geq Tr(\Sigma\Lambda) \ . \tag{2}$$

Lemma 3 *For any positive definite matrices $\tilde{M}, M$ with the same size, the following inequality holds when $0 < p \leq 2$:*

$$tr(\tilde{M}^{\frac{p}{2}}) - \frac{p}{2}tr(\tilde{M}M^{\frac{p-2}{2}}) \leq tr(M^{\frac{p}{2}}) - \frac{p}{2}tr(MM^{\frac{p-2}{2}}). \tag{3}$$

Proof. For any $\sigma > 0$, $\lambda > 0$ and $0 < p \leq 2$, according to Lemma 1 we have $\frac{p}{2}(\frac{\sigma}{\lambda}) - (\frac{\sigma}{\lambda})^{\frac{p}{2}} + \frac{2-p}{2} \geq 0$, which indicates

$$\frac{p}{2}\sigma\lambda^{\frac{p-2}{2}} - \sigma^{\frac{p}{2}} + \frac{2-p}{2}\lambda^{\frac{p}{2}} \geq 0 \ . \tag{4}$$

Suppose the eigen-decomposition $\tilde{M} = U\Sigma U^T$, $M = V\Lambda V^T$, where the eigenvalues in Σ is in increasing order and the eigenvalues in Λ is in decreasing order. Then according to Eq.(4), we have

$$\frac{p}{2}tr(\Sigma\Lambda^{\frac{p-2}{2}}) - tr(\Sigma^{\frac{p}{2}}) + \frac{2-p}{2}tr(\Lambda^{\frac{p}{2}}) \geq 0 \ , \tag{5}$$

and according to Lemma 2 we have

$$\frac{p}{2}tr(\tilde{M}M^{\frac{p-2}{2}}) - \frac{p}{2}tr(\Sigma\Lambda^{\frac{p-2}{2}}) \geq 0 \ , \tag{6}$$

$$\frac{p}{2}tr(\tilde{M}M^{\frac{p-2}{2}}) - tr(\Sigma^{\frac{p}{2}}) + \frac{2-p}{2}tr(\Lambda^{\frac{p}{2}}) \geq 0 \ . \tag{7}$$

Note that $tr(\tilde{M}^{\frac{p}{2}}) = tr(\Sigma^{\frac{p}{2}})$ and $tr(M^{\frac{p}{2}}) = tr(\Lambda^{\frac{p}{2}})$, so we have

$$\begin{aligned}
&\tfrac{p}{2}tr(\tilde{M}M^{\frac{p-2}{2}}) - tr(\tilde{M}^{\frac{p}{2}}) + \tfrac{2-p}{2}tr(M^{\frac{p}{2}}) \geq 0 \\
&\Rightarrow tr(\tilde{M}^{\frac{p}{2}}) - \tfrac{p}{2}tr(\tilde{M}M^{\frac{p-2}{2}}) \leq \tfrac{2-p}{2}tr(M^{\frac{p}{2}}) \\
&\Rightarrow tr(\tilde{M}^{\frac{p}{2}}) - \tfrac{p}{2}tr(\tilde{M}M^{\frac{p-2}{2}}) \leq tr(M^{\frac{p}{2}}) - \tfrac{p}{2}tr(MM^{\frac{p-2}{2}}) \ ,
\end{aligned}$$

which completes the proof. □

Lemma 4 *For any matrices $\tilde{A}, A$ with the same size and $\delta > 0$, the following inequality holds when $0 < p \leq 2$:*

$$\begin{aligned}
&tr((\tilde{A}^T\tilde{A} + \delta I)^{\frac{p}{2}}) - \tfrac{p}{2}tr(\tilde{A}^T\tilde{A}(A^TA + \delta I)^{\frac{p-2}{2}}) \\
&\leq tr((A^TA + \delta I)^{\frac{p}{2}}) - \tfrac{p}{2}tr(A^TA(A^TA + \delta I)^{\frac{p-2}{2}}) \ .
\end{aligned} \tag{8}$$

Proof. Note that $\tilde{A}^T\tilde{A} + \delta I$ and $A^TA + \delta I$ are positive definite matrices since $\delta > 0$. Then according to Lemma 3 we have

$$\begin{aligned}
&tr((\tilde{A}^T\tilde{A} + \delta I)^{\frac{p}{2}}) - \tfrac{p}{2}tr((\tilde{A}^T\tilde{A} + \delta I)(A^TA + \delta I)^{\frac{p-2}{2}}) \\
&\leq tr((A^TA + \delta I)^{\frac{p}{2}}) - \tfrac{p}{2}tr((A^TA + \delta I)(A^TA + \delta I)^{\frac{p-2}{2}}) \ ,
\end{aligned} \tag{9}$$

which indicates Eq.(8). □

As a result, we can prove Theorem **??** as follows now.

Proof of Theorem ??. In step 2 of Algorithm **??**, suppose the updated x is $\tilde{x}$. According to step 2, we know

$$f(\tilde{x}) + \sum_i tr(g_i^T(\tilde{x})g_i(\tilde{x})D_i) \leq f(x) + \sum_i tr(g_i^T(x)g_i(x)D_i) \ , \tag{10}$$

where the equality holds when and only when the algorithm converges.

For each i, according to Lemma 4, we have

$$\begin{array}{l} tr((g_i^T(\tilde{x})g_i(\tilde{x}) + \delta I)^{\frac{p}{2}}) \\ -\frac{p}{2}tr(g_i^T(\tilde{x})g_i(\tilde{x})(g_i^T(x)g_i(x) + \delta I)^{\frac{p-2}{2}}) \\ \leq tr((g_i^T(x)g_i(x) + \delta I)^{\frac{p}{2}}) \\ -\frac{p}{2}tr(g_i^T(x)g_i(x)(g_i^T(x)g_i(x) + \delta I)^{\frac{p-2}{2}}) \ . \end{array} \tag{11}$$

Note that $D_i = \frac{p}{2}(g_i^T(x)g_i(x) + \delta I)^{\frac{p-2}{2}}$, so for each i we have

$$\begin{array}{l} tr((g_i^T(\tilde{x})g_i(\tilde{x}) + \delta I)^{\frac{p}{2}}) - tr(g_i^T(\tilde{x})g_i(\tilde{x})D_i) \\ \leq tr((g_i^T(x)g_i(x) + \delta I)^{\frac{p}{2}}) - tr(g_i^T(x)g_i(x)D_i) \ . \end{array} \tag{12}$$

Then we have

$$\begin{array}{l} \sum_i tr((g_i^T(\tilde{x})g_i(\tilde{x}) + \delta I)^{\frac{p}{2}}) - \sum_i tr(g_i^T(\tilde{x})g_i(\tilde{x})D_i) \\ \leq \sum_i tr((g_i^T(x)g_i(x) + \delta I)^{\frac{p}{2}}) - \sum_i tr(g_i^T(x)g_i(x)D_i) \ . \end{array} \tag{13}$$

Summing Eq. (10) and Eq. (13) in the two sides, we arrive at

$$\begin{array}{c} f(\tilde{x}) + \sum_i tr((g_i^T(\tilde{x})g_i(\tilde{x}) + \delta I)^{\frac{p}{2}}) \leq \\ f(x) + \sum_i tr((g_i^T(x)g_i(x) + \delta I)^{\frac{p}{2}}) \ . \end{array} \tag{14}$$

Note that the equality in Eq.(14) holds only when the algorithm converges. Thus the Algorithm **??** will monotonically decrease the objective of the problem in Eq. (**??**) in each iteration until the algorithm converges. □

2 The Algorithm in the Method

Algorithm 1 The ADMM algorithm.

Set $1 < \rho < 2$ and initialize $\mu > 0$ and y.
while not converge **do**
1. Update x by solving $x^{k+1} = \arg\min_x (f(x) + \frac{\mu}{2}\|h(x, z^k) + \frac{y^k}{\mu}\|^2)$.
2. Update z by solving $z^{k+1} = \arg\min_z (g(z) + \frac{\mu}{2}\|h(x^{k+1}, z) + \frac{y^k}{\mu}\|^2)$
3. Update y by $y^{k+1} = y^k + \mu h(x^{k+1}, z^{k+1})$
4. Update μ by $\mu = \rho\mu$.
end while

3 The Detailed Procedures to Solve Eq. (**??**)

Step 1. Initialization.

Step 2. Solving $\mathbf{Y}$, when we fix the other variable $\mathbf{X}$ and the Lagrangian multiplier variable $\mathbf{\Lambda}$:

$$\min_{\mathbf{Y}} \ \mathbf{tr}\left(\mathbf{Y}^T\mathbf{L}\mathbf{X}\right) + \frac{\mu}{2}\left\|\mathbf{Y} - \mathbf{X} + \frac{1}{\mu}\mathbf{\Lambda}\right\|_{\mathrm{F}}^2, \quad s.t. \quad \mathbf{Y}^T\mathbf{Y} = \mathbf{I} \ . \tag{15}$$

Denoting $\mathbf{M} = \mu\mathbf{X} - \mathbf{\Lambda} - \mathbf{L}\mathbf{X}$, we can write the optimization problem in Eq. (15) as following:

$$\max_{\mathbf{Y}} \ \mathbf{tr}\left(\mathbf{Y}^T\mathbf{M}\right) \quad s.t. \quad \mathbf{Y}^T\mathbf{Y} = \mathbf{I} \ . \tag{16}$$

According to Theorem 1 in [**?**]the problem in Eq. (16) can be solved by computing the SVD of $\mathbf{M}$: if $svd\,(\mathbf{M}) = \mathbf{U}\mathbf{A}\mathbf{V}^T$, the solution of Eq. (16) is given by $\mathbf{U}\mathbf{V}^T$.

Step 3. Solving $\mathbf{X}$, when we fix the other variable $\mathbf{Y}$ and the Lagrangian multiplier variable $\mathbf{\Lambda}$:

$$\min_{\mathbf{X}} \ \mathbf{tr}\left(\mathbf{Y}^T\mathbf{L}\mathbf{X}\right) + \frac{\mu}{2}\left\|\mathbf{Y} - \mathbf{X} + \frac{1}{\mu}\mathbf{\Lambda}\right\|_{\mathrm{F}}^2, \quad s.t. \quad \mathbf{X} \geq 0 \ . \tag{17}$$

Denoting $\mathbf{N} = \mathbf{Y} + \frac{1}{\mu}\mathbf{\Lambda} - \frac{1}{\mu}\mathbf{L}^T\mathbf{Y}$, we can write the optimization problem in Eq. (17) as following:

$$\min_{\mathbf{X}} \ \|\mathbf{X} - \mathbf{N}\|_{\mathrm{F}}^2, \quad s.t. \quad \mathbf{X} \geq 0 \ , \tag{18}$$

which can be decoupled to solve the following problem for every entry of $\mathbf{X}$:

$$\min_{x_{ij}} \ (x_{ij} - n_{ij})^2 \ \ , \quad s.t. \quad x_{ij} \geq 0 \ , \tag{19}$$

which can be easily solved as follows: $x_{ij} = \max(n_{ij}, 0)$.

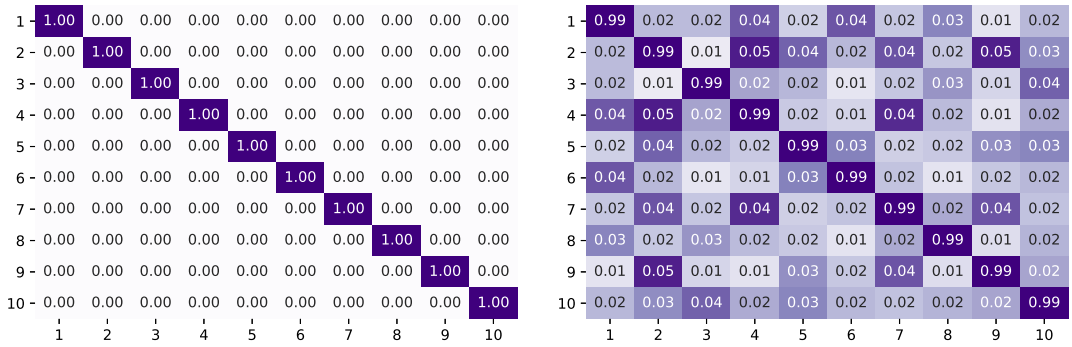
Step 4. Update $\mathbf{\Lambda}$ by $\mathbf{\Lambda} = \mathbf{\Lambda} + \mu(\mathbf{Y} - \mathbf{X})$.

Step 5. Update μ by $\mu = \rho\mu$.

4 Additional Experiments

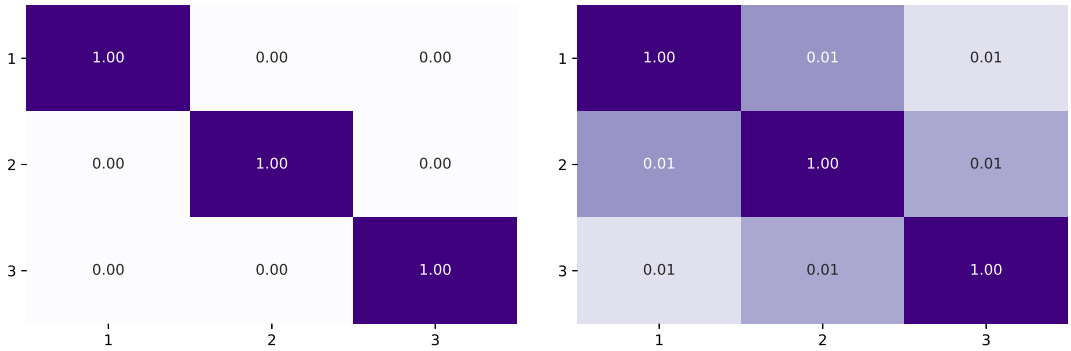
4.1 Orthogonality

Figure 1, 2, 3, 4 and 5 show the comparisons between our approach (on the left side) and NLE (on the right side). The result on AT&T data set could not be shown here due to large number of classes. In all 7 data sets, the solution of NLE method are very loosely constrained by $X^T X = I$. The experiment results so far show that our method can achieve a strict orthogonality for each data set which will in return have a better performance and robustness after embedding while NLE failed to guarantee the orthogonality constraint.



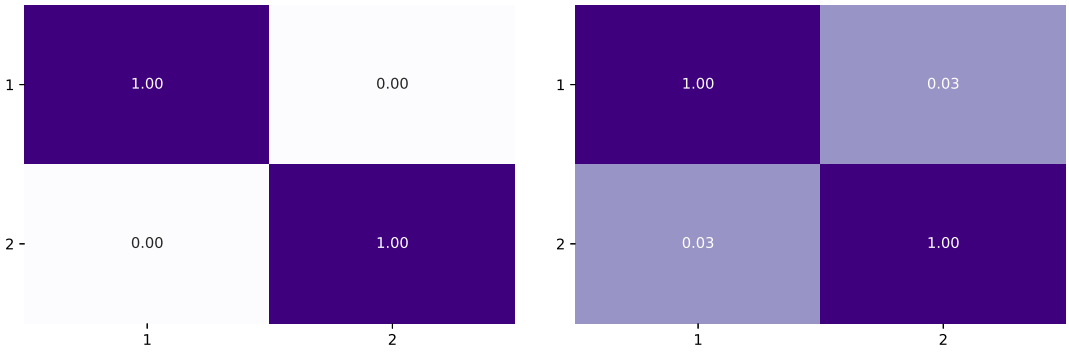
(a) Orthogonality of our method (b) Orthogonality of NLE method

Figure 1: The comparison of orthogonality between our method and NLE on minist data set.



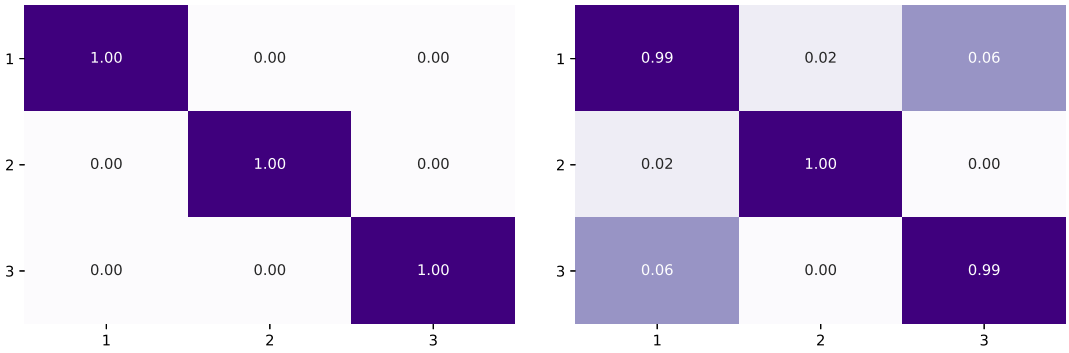
(a) Orthogonality of our method (b) Orthogonality of NLE method

Figure 2: The comparison of orthogonality between our method and NLE on caltech101 data set.



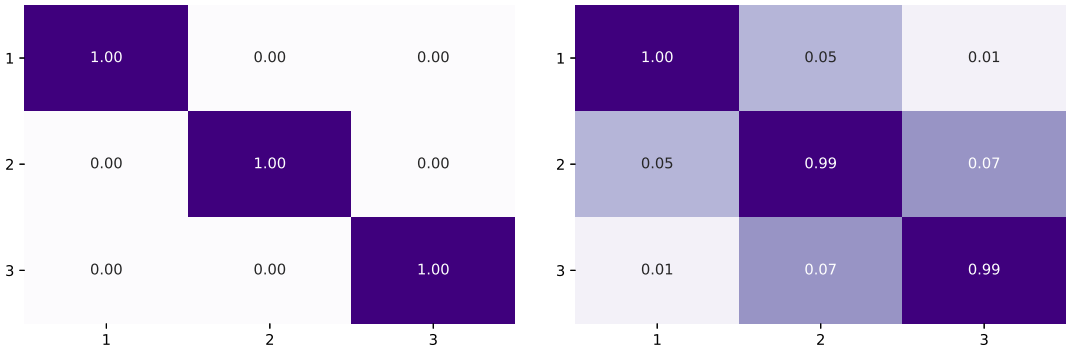
(a) Orthogonality of our method (b) Orthogonality of NLE method

Figure 3: The comparison of orthogonality between our method and NLE on ionosphere data set.



(a) Orthogonality of our method (b) Orthogonality of NLE method

Figure 4: The comparison of orthogonality between our method and NLE on wine data set.

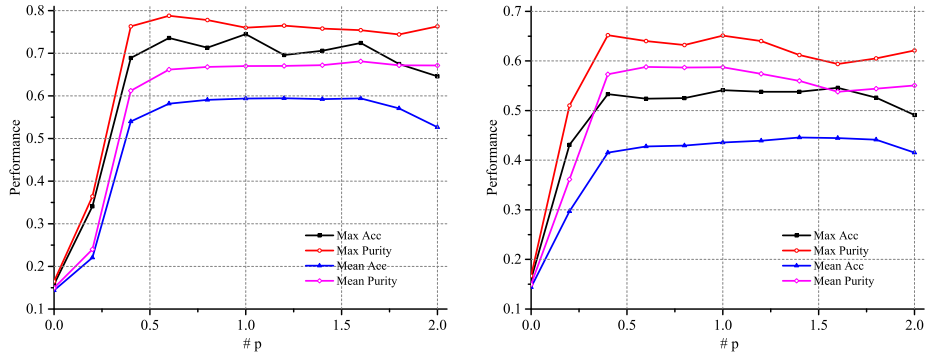


(a) Orthogonality of our method (b) Orthogonality of NLE method

Figure 5: The comparison of orthogonality between our method and NLE on iris data set.

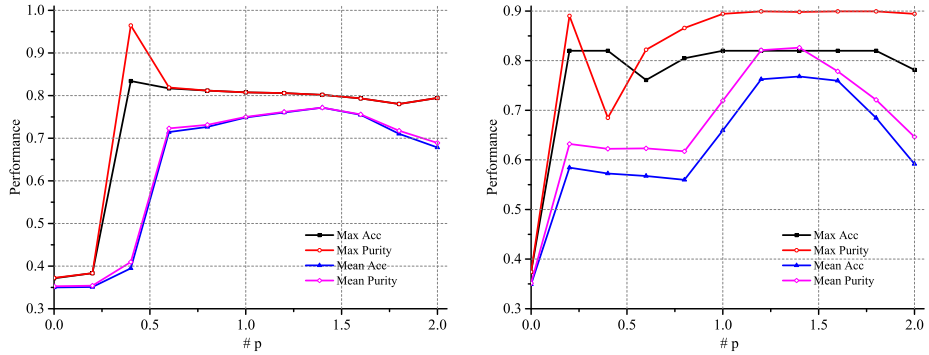
4.2 Performance on both Noiseless and Noisy Data

Figure 6, **??**, 7, 9, 8, 10 and 11 show the clustering performances of the proposed method on the experimental data sets as described in Section **??** and Section **??**.



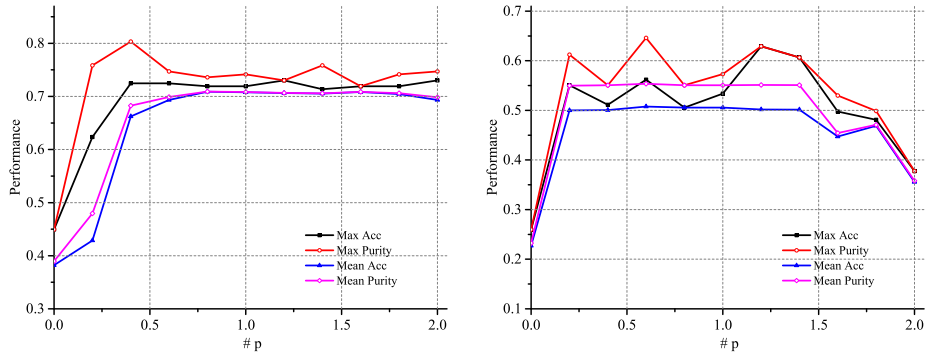
(a) Performance with original data (b) Performance with contaminated data

Figure 6: The comparison of performance on original and contaminated minist data set.



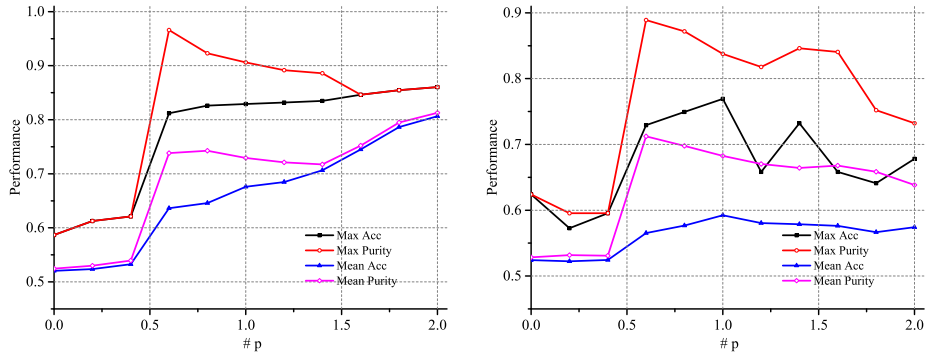
(a) Performance with original data (b) Performance with contaminated data

Figure 7: The comparison of performance on original and contaminated caltech101 data set.



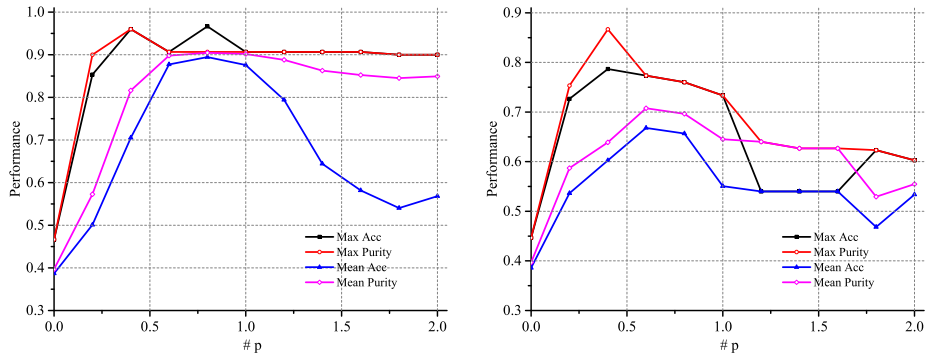
(a) Performance with original data (b) Performance with contaminated data

Figure 8: The comparison of performance on original and contaminated wine data set.



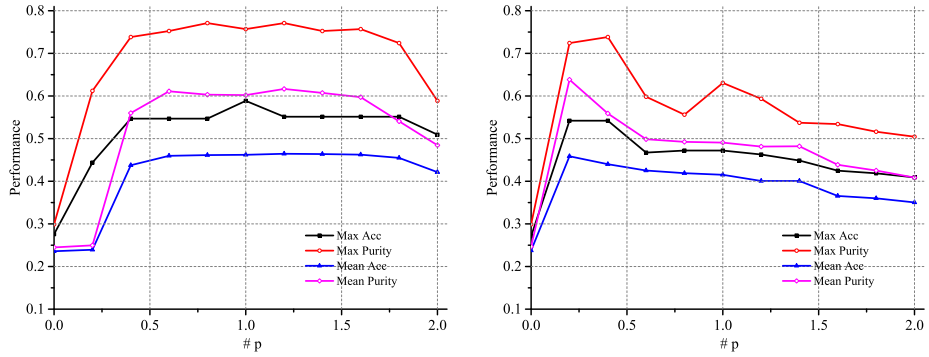
(a) Performance with original data (b) Performance with contaminated data

Figure 9: The comparison of performance on original and contaminated ionosphere data set.



(a) Performance with original data (b) Performance with contaminated data

Figure 10: The comparison of performance on original and contaminated iris data set.



(a) Performance with original data (b) Performance with contaminated data

Figure 11: The comparison of performance on original and contaminated glass data set.