Learning a Robust Consensus Matrix for Clustering Ensemble via Kullback-Leibler Divergence Minimization

Peng Zhou^{1,2} , **Liang Du**^{1,3*} , **Hanmo Wang**^{1,2} , **Lei Shi**^{1,2} and **Yi-Dong Shen**^{1*}
^{1}State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences
^{2}University of Chinese Academy of Sciences, Beijing 100049, China
^{3}School of Computer and Information Technology, Shanxi University, Taiyuan 030006, PR China
{zhoup,duliang,wanghm,shilei,ydshen}@ios.ac.cn

Abstract

Clustering ensemble has emerged as an important extension of the classical clustering problem. It provides a framework for combining multiple base clusterings of a data set to generate a final consensus result. Most existing clustering methods simply combine clustering results without taking into account the noises, which may degrade the clustering performance. In this paper, we propose a novel robust clustering ensemble method. To improve the robustness, we capture the sparse and symmetric errors and integrate them into our robust and consensus framework to learn a low-rank matrix. Since the optimization of the objective function is difficult to solve, we develop a block coordinate descent algorithm which is theoretically guaranteed to converge. Experimental results on real world data sets demonstrate the effectiveness of our method.

1 Introduction

Clustering is a fundamental problem in machine learning. According to [Wang *et al.*, 2009], traditional single clustering algorithms usually suffer from the robustness problems, because: (1) different clustering methods may discover very different structure in a given data set due to their different objective functions; (2) for a single clustering method, since no ground truth is available, we can hardly validate the clustering results; (3) some methods (such as K-means) are highly depend on their initializations. To improve the quality of clustering, the idea of ensemble has been proposed.

Clustering ensemble provides a framework for combining multiple base clusterings of a data set to generate a consensus clustering [Topchy *et al.*, 2004]. In past decades, many clustering ensemble methods have been proposed. [Strehl and Ghosh, 2003] formalized clustering ensemble as a combinatorial optimization problem in terms of shared mutual information. [Topchy *et al.*, 2003] also used information theoretic method to combine clusterings. [Fern and Brodley, 2004] proposed a graph cut method. [Li *et al.*, 2007; Li and Ding, 2008] and [Du *et al.*, 2011] applied non-negative matrix factorization (NMF) to clustering ensemble. [Wang *et al.*, 2009] learned the consensus clustering results by minimizing the Bregman divergence over all input clusterings. [Wang *et al.*, 2011] applied a Bayesian method to clustering ensemble. [Du *et al.*, 2013] proposed a self-supervised framework for clustering ensemble. These methods try to learn the consensus clustering results by taking advantage of diversity between base clusterings and reducing the redundance in clustering ensemble. In this paper we treat clustering ensemble with a new perspective. The inputs of clustering ensemble task are several weak base clustering results. Since the clusters identified by these base clusterings are often imperfect, these base clustering results may not be fully reliable. Thus these imperfect base clusterings can be regarded as intrinsic clustering corrupted with ”noises” and ”outliers”. As a result, it is necessary to recover these contaminated results for consensus clustering.

Most existing clustering ensemble methods blindly combine multiple base clusterings of data sets without taking into account noises and outliers, thus incurring the robustness problem, i.e., their performance would be severely degraded by noises and outliers. In this paper, we propose a novel Robust Clustering Ensemble (RCE) method, which explicitly characterizes the noises in each clustering, and uses them to get a robust and consensus clustering. In detail, given a set of input clusterings for ensemble we first construct the *connective matrices*, where each entry indicates the probability of two instances belonging to the same class. Since the input clusterings may contain noises and outliers, these connective matrices may also be contaminated. We introduce a sparse and symmetric error matrix for each connective matrix to explicitly identify the noises, and integrate the error matrices and connective matrices into ensemble framework. Considering that the connective matrices have a clear probabilistic interpretation, we use the Kullback-Leibler divergence for consensus measuring. We further impose a low-rank constraint on the final consensus matrix to obtain a more clear cluster structure. The resulting optimization problem turns out to be hard to solve due to the involvement of the noise matrices, divergence function and the low-rank constraint. To solve the objective function, we develop a block coordinate descent algorithm which can be theoretically guaranteed to converge.

Main contributions of our work are summarized as follows

- To improve the robustness of clustering ensemble, we introduce the sparse and symmetric error matrices to char-

[*] Corresponding author

acterize the noises in each input clustering. To learn the final robust, low-rank consensus matrix, we minimize the disagreements among the connective matrices using the Kullback-Leibler divergence.

- We propose a block coordinate descent algorithm to solve the complex objective function which involves Kullback-Leibler divergence, sparse term and low-rank term.
- The experiments on several benchmark data sets show that our method outperforms other compared algorithms, which indicates the importance of robustness for clustering ensemble.

2 Clustering Ensemble

Let $\mathcal{X} = \{x_1, x_2, ..., x_n\}$ be a set of n data points. Suppose we are given a set of m clusterings $\mathcal{C} = \{\mathcal{C}^1, \mathcal{C}^2, ..., \mathcal{C}^m\}$ of the data in $\mathcal{X}$, each clustering $\mathcal{C}^i$ consisting of a set of clusters $\{\pi_1^i, \pi_2^i, ..., \pi_k^i\}$, where k is the number of clusters and $\mathcal{X} = \cup_{j=1}^k \pi_j^i$. Note that the number of clusters k could be different for different clusterings.

From $\mathcal{C}$, we can construct symmetric *connective matrix* $\mathbf{A}^{(i)}$ for partition $\mathcal{C}^i$ as:

$$A_{pq}^{(i)} = \begin{cases} 1, & \text{if } x_p \text{ and } x_q \text{ belong to the same cluster,} \\ 0, & \text{otherwise.} \end{cases}$$

More generally, sometimes we can also get a *soft* connective matrix $\mathbf{A}^{(i)}$, where $A_{pq}^{(i)}$ denotes the possibility that x_p and x_q belong to the same cluster. Thus $A_{pq}^{(i)} \in [0, 1]$ for i from 1 to m. The task of clustering ensemble is to learn a *consensus matrix* $\tilde{\mathbf{A}}$ from $\mathbf{A}^{(1)}$,...,$\mathbf{A}^{(m)}$, where $\tilde{A}_{pq}$ denotes the *consensus* probability that x_p and x_q belong to the same cluster.

3 Robust Clustering Ensemble

In this section, we present the framework of our RCE method and then discuss how to solve it.

3.1 Formulation

Since $\mathbf{A}^{(1)}$,...,$\mathbf{A}^{(m)}$ may be contaminated by noises and outliers, we denote $\mathbf{E}^{(i)}$ as the sparse error matrix of the i-th connective matrix. The error matrix $\mathbf{E}^{(i)}$ should be symmetric because the connective matrix is symmetric. Given $\mathbf{E}^{(i)}$, we obtain a *cleaned* connective matrix $\mathbf{A}^{(i)} - \mathbf{E}^{(i)}$. To learn the consensus matrix, we minimize the average disagreement over the cleaned connective matrices. Since both $\tilde{A}_{pq}$ and $A_{pq}^{(i)} - E_{pq}^{(i)}$ have a probabilistic interpretation, we minimize the Kullback-Leibler divergence between $\tilde{A}_{pq}$ and $A_{pq}^{(i)} - E_{pq}^{(i)}$ instead of the Euclidean distance. Thus, we obtain the following optimization problem:

$$\min_{\tilde{\mathbf{A}}, \mathbf{E}^{(i)}} \sum_{i=1}^{m} \sum_{p,q=1}^{n} \mathrm{KL}(\tilde{A}_{pq}, A_{pq}^{(i)} - E_{pq}^{(i)}) + \lambda_1 \sum_{i=1}^{m} \|\mathbf{E}^{(i)}\|_1, \tag{1}$$

$$\text{s.t.} \quad \forall i, p, q, \quad 0 \leq \tilde{A}_{pq} \leq 1, \quad 0 \leq A_{pq}^{(i)} - E_{pq}^{(i)} \leq 1, \quad \mathbf{E}^{(i)} = \mathbf{E}^{(i)T}.$$

where $\mathrm{KL}(\cdot)$ is Kullback-Leibler divergence and λ_1 is a balancing parameter. $\|\cdot\|_1$ is the ℓ_1-norm, which makes $E^{(i)}$ sparse as [Du and Shen, 2013; He *et al.*, 2014] did. The constraints on $\tilde{A}_{pq}$ and $A_{pq}^{(i)} - E_{pq}^{(i)}$ ensure that the cleaned connective matrices and consensus matrix have a probabilistic interpretation. The constraint on $\mathbf{E}^{(i)}$ makes it symmetric.

Additionally, to make the final consensus matrix $\tilde{\mathbf{A}}$ have a clear structure for clustering, we further require $\tilde{\mathbf{A}}$ to be low-rank. Here we use the nuclear norm of $\tilde{\mathbf{A}}$ to approximate the rank of $\tilde{\mathbf{A}}$ inspired by [Liu *et al.*, 2010; Luo *et al.*, 2012; Liu *et al.*, 2013]:

$$\begin{aligned} \min_{\tilde{\mathbf{A}}, \mathbf{E}^{(i)}} \quad & \sum_{i=1}^{m} \sum_{p,q=1}^{n} \mathrm{KL}(\tilde{A}_{pq}, A_{pq}^{(i)} - E_{pq}^{(i)}) \\ & + \lambda_1 \sum_{i=1}^{m} \|\mathbf{E}^{(i)}\|_1 + \lambda_2 \|\tilde{\mathbf{A}}\|_*, \\ \text{s.t.} \quad & \forall i, p, q, \quad 0 \leq \tilde{A}_{pq} \leq 1, \quad 0 \leq A_{pq}^{(i)} - E_{pq}^{(i)} \leq 1, \qquad (2) \\ & \mathbf{E}^{(i)} = \mathbf{E}^{(i)T}. \end{aligned}$$

where λ_2 is another balancing parameter, and $\|\cdot\|_*$ is the nuclear norm.

Since $\tilde{A}_{pq}$ and $A_{pq}^{(i)} - E_{pq}^{(i)}$ are the possibility of x_p and x_q belonging to the same cluster, $\tilde{A}_{pq}$ and $A_{pq}^{(i)} - E_{pq}^{(i)}$ follow the Bernoulli distribution. Thus we can expand the $\mathrm{KL}(\cdot)$ and rewrite Eq.(2) to the following objective function (here 0log0 = 0):

$$\begin{aligned} \min_{\tilde{\mathbf{A}}, \mathbf{E}^{(i)}} \quad & \sum_{i=1}^{m} \sum_{p,q=1}^{n} \Bigg(\tilde{A}_{pq} \log \frac{\tilde{A}_{pq}}{A_{pq}^{(i)} - E_{pq}^{(i)}} \\ & + (1 - \tilde{A}_{pq}) \log \frac{1 - \tilde{A}_{pq}}{1 - (A_{pq}^{(i)} - E_{pq}^{(i)})} \Bigg) \\ & + \lambda_1 \sum_{i=1}^{m} \|\mathbf{E}^{(i)}\|_1 + \lambda_2 \|\tilde{\mathbf{A}}\|_*, \\ \text{s.t.} \quad & \forall i, p, q, \quad 0 \leq \tilde{A}_{pq} \leq 1, \quad 0 \leq A_{pq}^{(i)} - E_{pq}^{(i)} \leq 1, \\ & \mathbf{E}^{(i)} = \mathbf{E}^{(i)T}. \end{aligned} \tag{3}$$

Observe that due to the explicit characterization ($\mathbf{E}^{(i)}$) of sparse, symmetric and bounded noises in clusterings, the above framework of clustering ensemble is robust. Moreover, to be more suitable for clustering ensemble task, we also introduce the Kullback-Leibler divergence and low-rank term in our approach.

3.2 Optimization

Eq.(3) involves two groups variables ($\tilde{\mathbf{A}}$ and $\mathbf{E}^{(i)}$), thus we present a block coordinate descent scheme to optimize it. In particular, we optimize the objective with respect to one variable while fixing the other variables. This procedure repeats until convergence.

To handle the nuclear norm term $\|\tilde{\mathbf{A}}\|_*$, we borrow the following result from [Grave *et al.*, 2011]:

Lemma 1. *Let* $\mathbf{M} \in \mathcal{R}^{n\times m}$*. The nuclear norm of* $\mathbf{M}$ *is equal to:*

$$\|\mathbf{M}\|_* = \frac{1}{2}\left(\inf_{\mathbf{S}\succeq 0} tr(\mathbf{M}^T\mathbf{S}^{-1}\mathbf{M}) + tr(\mathbf{S})\right)$$

and the infimum is attained for $\mathbf{S} = (\mathbf{M}\mathbf{M}^T)^{1/2}$*, where* $tr(\cdot)$ *is the trace of a matrix, and* $\mathbf{S} \succeq 0$ *means* $\mathbf{S}$ *is positive semi-definite.*

According to Lemma 1, we rewrite Eq.(3) as:

$$\begin{aligned}\min_{\tilde{\mathbf{A}},\mathbf{E}^{(i)},\mathbf{S}} \quad & \sum_{i=1}^{m}\sum_{p,q=1}^{n}\left(\tilde{A}_{pq}\log\frac{\tilde{A}_{pq}}{A_{pq}^{(i)} - E_{pq}^{(i)}}\right.\\ & \left. + (1-\tilde{A}_{pq})\log\frac{1-\tilde{A}_{pq}}{1-(A_{pq}^{(i)} - E_{pq}^{(i)})}\right)\\ & + \lambda_1\sum_{i=1}^{m}\|\mathbf{E}^{(i)}\|_1 + \lambda_2\left(\mathrm{tr}(\tilde{\mathbf{A}}^T\mathbf{S}^{-1}\tilde{\mathbf{A}}) + \mathrm{tr}(\mathbf{S})\right),\\ \text{s.t.} \quad & \forall p,q,i, \quad 0 \le \tilde{A}_{pq} \le 1, \quad 0 \le A_{pq}^{(i)} - E_{pq}^{(i)} \le 1,\\ & \mathbf{E}^{(i)} = \mathbf{E}^{(i)T}, \quad \mathbf{S} \succeq 0. \end{aligned} \tag{4}$$

Optimize S by fixing $\tilde{\mathbf{A}}$ and $\mathbf{E}^{(i)}$

When fixing $\tilde{\mathbf{A}}$ and $\mathbf{E}^{(i)}$, Eq.(4) is re-written as:

$$\begin{aligned}\min_{\mathbf{S}} \quad & \mathrm{tr}(\tilde{\mathbf{A}}^T\mathbf{S}^{-1}\tilde{\mathbf{A}}) + \mathrm{tr}(\mathbf{S}),\\ \text{s.t.} \quad & \mathbf{S} \succeq 0.\end{aligned} \tag{5}$$

As suggested by [Grave *et al.*, 2011], we need to add a term $\mu\mathrm{tr}(\mathbf{S}^{-1})$. Otherwise, the infimum over $\mathbf{S}$ could be attained at a non-invertible $\mathbf{S}$, leading to a non-convergent algorithm. Here μ is a small parameter and fixed to 0.001 for simplicity. According to Lemma 1, the infimum over $\mathbf{S}$ is then attained for:

$$\mathbf{S} = (\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T + \mu\mathbf{I})^{1/2}. \tag{6}$$

where $\mathbf{I}$ is an identity matrix. Considering that when updating other variables, we just need $\mathbf{S}^{-1}$ instead of $\mathbf{S}$, we calculate $\mathbf{S}^{-1}$ directly. More specifically, we compute $\mathbf{V}Diag(\sigma_k)\mathbf{V}^T$ as the eigenvalue decomposition of $\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T$, and then calculate $\mathbf{S}^{-1}$ as $\mathbf{S}^{-1} = \mathbf{V}Diag(1/\sqrt{\sigma_k + \mu})\mathbf{V}^T$.

Optimize $\tilde{\mathbf{A}}$ by fixing S and $\mathbf{E}^{(i)}$

By fixing $\mathbf{E}^{(i)}$ and $\mathbf{S}$, Eq.(4) can be simplified as:

$$\begin{aligned}\min_{\tilde{\mathbf{A}}} \quad & \sum_{i=1}^{m}\sum_{p,q=1}^{n}\left(\tilde{A}_{pq}\log\frac{\tilde{A}_{pq}}{A_{pq}^{(i)} - E_{pq}^{(i)}}\right.\\ & \left. + (1-\tilde{A}_{pq})\log\frac{1-\tilde{A}_{pq}}{1-(A_{pq}^{(i)} - E_{pq}^{(i)})}\right)\\ & + \lambda_2\mathrm{tr}(\tilde{\mathbf{A}}^T\mathbf{S}^{-1}\tilde{\mathbf{A}}),\\ \text{s.t.} \quad & \forall p,q, \quad 0 \le \tilde{A}_{pq} \le 1.\end{aligned} \tag{7}$$

For the constraint $0 \le \tilde{A}_{pq} \le 1$, we cannot get the closed-form of $\tilde{\mathbf{A}}$. To solve Eq.(7), we use the auxiliary function approach [Lee and Seung, 2000]. We first introduce the definition and lemma of auxiliary function.

Definition 1. *[Lee and Seung, 2000]* $Z(h, h')$ *is an auxiliary function for* $F(h)$ *if the conditions*

$$Z(h, h') \ge F(h), Z(h, h) = F(h),$$

are satisfied.

Lemma 2. *[Lee and Seung, 2000] If* Z *is an auxiliary function for* F*, the* F *is non-increasing under the update*

$$h^{(t+1)} = \arg\min_h Z(h, h^{(t)}) \tag{8}$$

To find an auxiliary function for Eq.(7), we first denote

$$\begin{aligned} C_{pq} &= \sum_{i=1}^{m}\log\frac{1 - A_{pq}^{(i)} + E_{pq}^{(i)}}{A_{pq}^{(i)} - E_{pq}^{(i)}},\\ \mathbf{D} &= \lambda_2\mathbf{S}^{-1}. \end{aligned}$$

$\mathbf{C}$ is a matrix whose (p,q)-th element is C_{pq}. We further introduce $\mathbf{C}^+ = (|\mathbf{C}| + \mathbf{C})/2$, $\mathbf{C}^- = (|\mathbf{C}| - \mathbf{C})/2$, $\mathbf{D}^+ = (|\mathbf{D}| + \mathbf{D})/2$ and $\mathbf{D}^- = (|\mathbf{D}| - \mathbf{D})/2$. Then following Theorem defines an auxiliary function of Eq.(7).

Theorem 1. *Let* $J(\tilde{\mathbf{A}})$ *be the objective function of Eq.(7), then the following function*

$$\begin{aligned} & Z(\tilde{\mathbf{A}}, \mathbf{A}') \\ = & \sum_{p,q=1}^{n} m\left(\frac{\tilde{A}_{pq}^2}{A'_{pq}} + \tilde{A}_{pq}\log A'_{pq} + (1-\tilde{A}_{pq})\log(1-A'_{pq}) - 1\right.\\ & \left. + \frac{(1-\tilde{A}_{pq})^2}{1-A'_{pq}}\right) - \sum_{p,q=1}^{n}\left(C_{pq}^- A'_{pq}\left(1 + \log\frac{\tilde{A}_{pq}}{A'_{pq}}\right)\right)\\ & + \sum_{p,q=1}^{n} C_{pq}^+\frac{\tilde{A}_{pq}^2 + A'^2_{pq}}{2A'_{pq}} + \sum_{p,q=1}^{n}\frac{(\mathbf{D}^+\mathbf{A}')_{pq}\tilde{A}_{pq}^2}{A'_{pq}}\\ & - \sum_{p,q,r=1}^{n} D_{qr}^- A'_{qp}A'_{rp}\left(1 + \log\frac{\tilde{A}_{qp}\tilde{A}_{rp}}{A'_{qp}A'_{rp}}\right) \end{aligned} \tag{9}$$

is an auxiliary function of $J(\tilde{\mathbf{A}})$.

Proof. See Appendix A in the supplementary material. □

According to Lemma 2, we minimize $Z(\tilde{\mathbf{A}}, \mathbf{A}')$ instead of $J(\tilde{\mathbf{A}})$. Let $\frac{\partial Z(\tilde{\mathbf{A}},\mathbf{A}')}{\partial\tilde{\mathbf{A}}} = 0$. We get the updating rule of $\tilde{\mathbf{A}}$:

$$\tilde{A}_{pq}^{(t+1)} \leftarrow \frac{-B_{pq}^{(t)} + \sqrt{B_{pq}^{(t)2} + 4G_{pq}^{(t)}H_{pq}^{(t)}}}{2G_{pq}^{(t)}} \tag{10}$$

where $\tilde{A}_{pq}^{(t+1)}$ expresses the value of $\tilde{A}_{pq}$ in the $(t+1)$-th iteration, and

$$\begin{aligned} G_{pq}^{(t)} &= \frac{2m}{\tilde{A}_{pq}^{(t)}} + \frac{2m}{1-\tilde{A}_{pq}^{(t)}} + \frac{C_{pq}^+}{\tilde{A}_{pq}^{(t)}} + \frac{2(\mathbf{D}^+\tilde{\mathbf{A}}^{(t)})_{pq}}{\tilde{A}_{pq}^{(t)}}\\ B_{pq}^{(t)} &= m\log\tilde{A}_{pq}^{(t)} - \frac{2m}{1-\tilde{A}_{pq}^{(t)}} - m\log(1-\tilde{A}_{pq}^{(t)})\\ H_{pq}^{(t)} &= C_{pq}^-\tilde{A}_{pq}^{(t)} + 2(\mathbf{D}^-\tilde{\mathbf{A}}^{(t)})_{pq}\tilde{A}_{pq}^{(t)} \end{aligned}$$

It is obvious that $\tilde{A}_{pq} \geq 0$ always holds when updating $\tilde{A}_{pq}$ by Eq.(10). When $\tilde{A}_{pq}$ is greater than 1, since $Z(\tilde{\mathbf{A}}, \mathbf{A}')$ is convex w.r.t. $\tilde{\mathbf{A}}$, $Z(\tilde{\mathbf{A}}, \mathbf{A}')$ is a monotone decreasing function in interval $[0, 1]$. Thus the optima of $\tilde{A}_{pq}$ is at $\tilde{A}_{pq} = 1$. To sum up, we update $\tilde{A}_{pq}$ as follows:

$$\tilde{A}^{(t)}_{pq} \leftarrow \min\left\{\frac{-B^{(t)}_{pq}+\sqrt{B^{(t)2}_{pq}+4G^{(t)}_{pq}H^{(t)}_{pq}}}{2G^{(t)}_{pq}}, 1\right\} \tag{11}$$

Theorem 2. *Updating $\tilde{\mathbf{A}}$ by Eq.(11) will monotonically decrease the value of the objective of Eq.(7); hence it converges.*

Proof. By Lemma 2 and Theorem 1, we can get that $J(\tilde{\mathbf{A}}^{(1)}) = Z(\tilde{\mathbf{A}}^{(1)}, \tilde{\mathbf{A}}^{(1)}) \geq Z(\tilde{\mathbf{A}}^{(2)}, \tilde{\mathbf{A}}^{(1)}) \geq J(\tilde{\mathbf{A}}^{(2)}) \geq$ .... So $J(\tilde{\mathbf{A}})$ is monotonically decreasing. Since $J(\tilde{\mathbf{A}})$ is obviously bounded below, we prove this theorem. □

Optimize $\mathbf{E}^{(i)}$ by fixing $\mathbf{S}$ and $\tilde{\mathbf{A}}$

When $\mathbf{S}$ and $\tilde{\mathbf{A}}$ are fixed, the remaining problem can be further decomposed into $n \times n \times m$ sub-problems, where only $E^{(i)}_{pq}$ is involved. Considering the symmetry of $\mathbf{A}^{(i)}$ and $\mathbf{E}^{(i)}$, there are two cases, i.e. off-diagonal and diagonal. For the i-th connective matrix, the sub-problem of the off-diagonal element $E^{(i)}_{pq}$ $(p \neq q)$ is:

$$\begin{aligned}\min_{E^{(i)}_{pq}} & -\tilde{A}_{pq}\log(A^{(i)}_{pq}-E^{(i)}_{pq})-(1-\tilde{A}_{pq})\log(1-A^{(i)}_{pq}+E^{(i)}_{pq}) \\ & -\tilde{A}_{qp}\log(A^{(i)}_{pq}-E^{(i)}_{pq})-(1-\tilde{A}_{qp})\log(1-A^{(i)}_{pq}+E^{(i)}_{pq}) \\ & +2\lambda_1|E^{(i)}_{pq}|, \qquad\qquad (12) \\ \text{s.t.} & A^{(i)}_{pq}-1 \leq E^{(i)}_{pq} \leq A^{(i)}_{pq}.\end{aligned}$$

The sub-problem of the diagonal element $E^{(i)}_{jj}$ is similar:

$$\begin{aligned}\min_{E^{(i)}_{jj}} & -\tilde{A}_{jj}\log(A^{(i)}_{jj}-E^{(i)}_{jj})-(1-\tilde{A}_{jj})\log(1-A^{(i)}_{jj}+E^{(i)}_{jj}) \\ & +\lambda_1|E^{(i)}_{jj}|, \qquad\qquad (13) \\ \text{s.t.} & A^{(i)}_{jj}-1 \leq E^{(i)}_{jj} \leq A^{(i)}_{jj}.\end{aligned}$$

Since Eq.(13) can be regarded as half of Eq.(12), we only need to solve Eq.(12); and the solution of Eq.(13) is similar.

For simplicity, we denote $a = \tilde{A}_{pq}$, $b = \tilde{A}_{qp}$, $c = A^{(i)}_{pq}$ and $x = E^{(i)}_{pq}$. Theorem 3 gives the solution to Eq.(12).

Theorem 3. *Eq.(12) is equivalent to*

$$\begin{aligned}\min_{x} \quad & f(x) = -a\log(c-x)-(1-a)\log(1-c+x) \\ & -b\log(c-x)-(1-b)\log(1-c+x)+2\lambda_1|x|, \\ s.t. \quad & c-1 \leq x \leq c. \qquad\qquad (14)\end{aligned}$$

and the solution of Eq.(14) is

$$x = \arg\min_{x_1,x_2}\{f(x_1), f(x_2)\}.$$

where

$$x_1 = \min\left\{\frac{2\lambda_1 c-\lambda_1-1+\sqrt{(\lambda_1+1)^2-2\lambda_1(a+b)}}{2\lambda_1}, 0\right\},$$

$$x_2 = \max\left\{\frac{\lambda_1-2\lambda_1 c-1+\sqrt{(1-\lambda_1)^2+2\lambda_1(a+b)}}{-2\lambda_1}, 0\right\}.$$

Proof. It is obvious that Eq.(14) is equal to Eq.(12). Now we solve Eq.(14). Since there is an absolute value function in Eq.(14), we consider two cases: $x < 0$ and $x \geq 0$.

When $x < 0$, setting the derivative of Eq.(14) w.r.t. x to zero, we get:

$$\frac{a}{c-x}-\frac{1-a}{1-c+x}+\frac{b}{c-x}-\frac{1-b}{1-c+x}-2\lambda_1 = 0 \tag{15}$$

Solve x from Eq.(15):

$$x = \frac{2\lambda_1 c-\lambda_1-1 \pm \sqrt{(\lambda_1+1)^2-2\lambda_1(a+b)}}{2\lambda_1} \tag{16}$$

To handle the constraint in Eq.(14), we introduce Lemma 3:

Lemma 3.

$$\frac{2\lambda_1 c-\lambda_1-1-\sqrt{(\lambda_1+1)^2-2\lambda_1(a+b)}}{2\lambda_1} \leq c-1$$

and

$$c-1 \leq \frac{2\lambda_1 c-\lambda_1-1+\sqrt{(\lambda_1+1)^2-2\lambda_1(a+b)}}{2\lambda_1} \leq c$$

satisfies at any value of $\lambda_1 > 0$ when $a, b \in [0, 1]$.

Proof. See Appendix B in the supplementary material. □

According to Lemma 3, to satisfy the constraint in Eq.(14), we can only take the second solution. Although the first solution sometimes can reach the boundary value $c-1$, it is easy to verify that when $x = c-1$, the objective function becomes infinitely large. Moreover, if $\frac{2\lambda_1 c-\lambda_1-1+\sqrt{(\lambda_1+1)^2-2\lambda_1(a+b)}}{2\lambda_1} > 0$, which means the objective function is monotonically decreasing in $[c-1, 0]$, thus the optima is 0. To sum up, if $x < 0$, then the optima is:

$$x = \min\left\{\frac{2\lambda_1 c-\lambda_1-1+\sqrt{(\lambda_1+1)^2-2\lambda_1(a+b)}}{2\lambda_1}, 0\right\}. \tag{17}$$

Similarly, when $x \geq 0$, the optima is:

$$x = \max\left\{\frac{\lambda_1-2\lambda_1 c-1+\sqrt{(1-\lambda_1)^2+2\lambda_1(a+b)}}{-2\lambda_1}, 0\right\}. \tag{18}$$

Therefore, to optimize x, we compute two candidate values according to Eq.(17) and Eq.(18), then we choose the one which makes objective function of Eq.(14) smaller. □

Algorithm 1 summarizes the whole optimization process. After getting $\tilde{\mathbf{A}}$, we set $\mathbf{W} = (\tilde{\mathbf{A}}+\tilde{\mathbf{A}}^T)/2$, and apply spectral clustering [Ng *et al.*, 2001] on $\mathbf{W}$ to get the final clustering.

Algorithm 1 Robust Clustering Ensemble

```
Input: multiple connective matrices $\mathbf{A}^{(1)}, ..., \mathbf{A}^{(m)}$, parameters $\lambda_1$, $\lambda_2$ and $\mu$.
Output: consensus matrix $\tilde{\mathbf{A}}$, error matrices $\mathbf{E}^{(i)}$.
Initialize $\tilde{\mathbf{A}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{A}^{(i)}$.
while not converge do
    Compute $\mathbf{E}^{(i)}$ according to Theorem 3.
    Compute $\mathbf{V}Diag(\sigma_k)\mathbf{V}^T$ as the eigenvalue decomposition of $\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T$.
    Set $\mathbf{S}^{-1} = \mathbf{V}Diag(1/\sqrt{\sigma_k + \mu})\mathbf{V}^T$.
    Update $\tilde{\mathbf{A}}$ by Eq.(11).
end while
```

3.3 Convergence and Complexity Analysis

According to Theorem 2, when updating $\tilde{\mathbf{A}}$, the objective function decreases monotonically. When updating $\mathbf{E}^{(i)}$, we find the global solution of this sub-problem which also makes the objective function decrease. In addition, the objective function has a lower bound. Thus Algorithm 1 converges.

In each iteration, when updating $\mathbf{S}^{-1}$, the time complexity is $O(n^3)$ due to the eigenvalue decomposition, where n is the number of instances. When updating $\tilde{\mathbf{A}}$, since there is a matrix multiplication of two $n \times n$ matrices ($\mathbf{D}^{+}\tilde{\mathbf{A}}$), the complexity is $O(n^3)$. When computing $E^{(i)}$, which only contains element-wise operation, the complexity is $O(n^2 m)$, where m is the number of connective matrices. Thus the overall complexity is $O((n^3 + n^2 m)l)$, where l is the number of iterations.

4 Experiments

In this section, we evaluate the effectiveness of the proposed RCE method by comparing with several state-of-the-art clustering ensemble methods on benchmark data sets.

4.1 Data Sets

We use totally 11 data sets to evaluate the effectiveness of our proposed RCE, including images, texts, and UCI data sets. Data sets from different areas serve as a good test bed for a comprehensive evaluation. The basic information of these data sets are summarized in Table 1.

Table 1: Description of the data sets.

	#instances	#features	#classes
Yale	165	1024	15
Tr23	204	5832	6
JAFFE	213	676	10
Glass	214	9	6
ORL	400	1024	40
Medical	706	1449	17
Coil20	1440	1024	20
Wap	1560	8460	20
Hitech	2301	22498	6
K1b	2340	21839	6
MNIST	4000	784	10

4.2 Compared Methods

we compare RCE with the following algorithms:

- **K-means**, which is randomly initialized and the results are averaged over 200 independent runs.
- **KC**, which represents the results of applying K-means to a consensus similarity matrix, which is often used as a baseline in clustering ensemble methods such as [Li and Ding, 2008].
- **Cluster-based Similarity Partitioning Algorithm (CSPA)**[Strehl and Ghosh, 2003], which signifies a relationship between objects in the same cluster and can thus be used to establish a measure of pairwise similarity.
- **HyperGraph Partitioning Algorithm (HGPA)**[Strehl and Ghosh, 2003], which approximates the maximum mutual information objective with a constrained minimum cut objective.
- **Meta-CLustering Algorithm (MCLA)**[Strehl and Ghosh, 2003], the objective of integration is viewed as a cluster correspondence problem.
- **Nonnegative Matrix Factorization based Consensus clustering (NMFC)**[Li and Ding, 2008], which uses NMF to aggregate clustering results.
- **Generalized Weighted Cluster Aggregation (GWCA)**[Wang *et al.*, 2009], which learns the consensus clustering results by minimizing the Bregman divergence over all the input clusterings.
- **Bayesian Clustering Ensemble (BCE)**[Wang *et al.*, 2011], which is a Bayesian model for ensemble.
- **Robust Multiview Spectral Clustering (RMSC)**[Xia *et al.*, 2014]. Although it is not designed for clustering ensemble task, it also contains mechanism to handle noises and can be easily adopted to deal with clustering ensemble. We conduct the probabilistic transition matrices with the connective matrices, and then apply RMSC on them to obtain final clustering results.

4.3 Experiment Setup

Following the similar experimental protocol in [Wang *et al.*, 2011], we run k-means 200 times with different initializations to obtain 200 base clustering results, which are divided evenly into 10 subsets, with 20 base results in each of them. Clustering ensemble methods are then applied on each subset.

The number of clusters is set to be the true number of classes for all data sets and algorithms. We fix the parameter μ to 0.001 when updating $\mathbf{S}$ and tune both the parameters λ_1 and λ_2 from $[10^{-4}, 10^{4}]$ by grid search. We tune the parameters in compared methods as suggested in their papers. To measure the performance of clustering, the average clustering Accuracy (ACC) and Normalized Mutual Information (NMI) on the 10 subsets are reported. To validate the statistic significance of results, we also calculate the p-value of t-test.

4.4 Experimental Results

Table 2 shows the clustering results. Bold font expresses that the difference is statistically significant (p-value of t-test is

Table 2: Clustering Ensemble Results

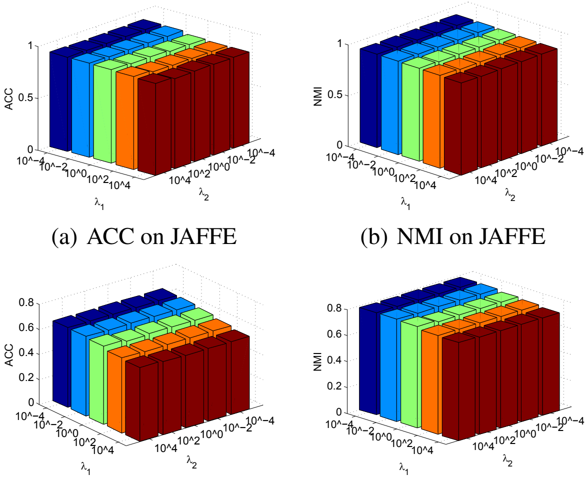
Dataset	Metric	K-means	KC	CSPA	HGPA	MCLA	NMFC	GWCA	BCE	RMSC	RCE
Yale	ACC	0.3673	0.3836	0.4079	0.3952	0.4121	0.3897	0.3836	0.4121	0.4186	**0.4332**
	NMI	0.4175	0.4254	0.4591	0.4522	0.4439	0.4411	0.4237	0.4505	0.4623	**0.4807**
Tr23	ACC	0.3904	0.3588	0.2946	0.3289	0.3363	0.3770	**0.4113**	0.3843	0.3786	**0.4163**
	NMI	0.1351	0.1394	0.0991	0.1357	0.1196	0.1393	0.1357	0.1554	0.1580	**0.1718**
JAFFE	ACC	0.7235	0.7117	0.8291	0.8601	0.8854	0.7127	0.7930	0.8042	0.7642	**0.9115**
	NMI	0.8098	0.8092	0.8351	0.8628	0.8979	0.8064	0.8561	0.8645	0.8365	**0.9221**
Glass	ACC	0.5086	0.4804	0.4164	0.3762	0.4967	0.4636	**0.5308**	0.4799	0.4848	**0.5292**
	NMI	0.3572	0.3456	0.2789	0.2102	0.3456	0.3427	**0.3755**	0.3435	0.3434	**0.3880**
ORL	ACC	0.5243	0.5528	0.6050	0.6203	0.6130	0.5760	0.5703	0.4285	0.5594	**0.6456**
	NMI	0.7278	0.7492	0.7728	**0.7838**	0.7740	0.7661	0.7558	0.5409	0.7499	**0.7922**
Medical	ACC	0.3959	0.3703	0.3504	0.3181	0.4021	0.3739	0.3950	0.3929	0.3831	**0.4220**
	NMI	0.4128	0.4161	0.3971	0.3844	0.4209	0.4278	0.4172	0.4436	0.4383	**0.4611**
Coil20	ACC	0.5931	0.6260	0.6744	0.5653	0.7166	0.6167	0.6354	0.6630	0.6180	**0.7263**
	NMI	0.7390	0.7530	0.7538	0.6915	0.7945	0.7503	0.7637	0.7757	0.7446	**0.8090**
Wap	ACC	0.3621	0.3404	0.2502	0.2473	0.2255	0.3601	0.3695	0.3654	0.3529	**0.4214**
	NMI	0.2866	0.3731	0.3300	0.2409	0.0274	0.3747	0.3107	0.3432	0.4003	**0.4251**
Hitech	ACC	0.3158	0.3157	0.3018	0.2402	0.3094	0.3130	0.3212	**0.3299**	0.3161	**0.3363**
	NMI	0.0967	0.1248	0.1226	0.0318	0.1035	0.1158	0.0942	0.1319	0.1495	**0.1656**
K1b	ACC	0.6794	0.6904	0.4267	0.3209	0.5302	0.6822	0.7241	**0.7429**	0.6526	**0.7524**
	NMI	0.2440	0.4266	0.3126	0.1161	0.2716	0.4152	0.2862	0.4235	0.4343	**0.4887**
MNIST	ACC	0.4948	0.5159	0.5097	0.4027	0.5252	0.4792	0.5161	0.5162	0.5136	**0.5302**
	NMI	0.4586	0.4555	0.4392	0.3909	0.4659	0.4485	0.4485	0.4709	0.4618	**0.4876**

less than 0.05). From Table 2, it can be seen that most clustering ensemble methods perform better than K-means, which indicates the benefit of ensemble methods. It can also be seen that, our proposed method RCE shows superior performance gains over the baselines w.r.t. ACC and NMI on most of the 11 data sets, which demonstrates that our robust ensemble can improve the performance of clustering.

RMSC was introduced for multivew spectral clustering. It has a mechanism to handle noises in multiple views, but differs significantly from ours. First, it is based on probabilistic transition matrix instead of connective matrix; second, it considers only a sparse error matrix, while we consider a symmetric, sparse and bounded matrix; it seems that our characterization of noises is more suitable for ensemble clustering tasks, as demonstrated in our experiments.

4.5 Parameter Study

We explore the effect of the parameters on clustering performance. There are two parameters in our method: λ_1 and λ_2. We tune these two parameters from $[10^{-4}, 10^4]$. We show the ACC and NMI on JAFFE and ORL data sets and the results are similar on other data sets. Figure 1 shows the results, from which we see that the performance of our method is stable across a wide range of the parameters.



(a) ACC on JAFFE (b) NMI on JAFFE

(c) ACC on ORL (d) NMI on ORL

Figure 1: ACC and NMI w.r.t λ_1, λ_2 on JAFFE and ORL.

5 Conclusion

In this paper we proposed a unified framework for robust clustering ensemble. We introduced symmetric and sparse error matrices to characterize noises and integrated them into a robust framework to learn a low-rank consensus matrix. We presented a block coordinate descent algorithm to solve the induced hard optimization problem and proved its convergence. Finally, experiments on benchmark data sets demonstrated the effectiveness of our method.

As ongoing work we are considering methods for handling connective matrices with missing values.

Acknowledgments

This work is supported in part by China National 973 program 2014CB340301 and NSFC grant 61379043, 61273291.

References

[Du and Shen, 2013] Liang Du and Yi-Dong Shen. Towards robust co-clustering. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 1317–1322. AAAI Press, 2013.

[Du *et al.*, 2011] Liang Du, Xuan Li, and Yi-Dong Shen. Cluster ensembles via weighted graph regularized nonnegative matrix factorization. In *Advanced Data Mining and Applications*, pages 215–228. Springer, 2011.

[Du *et al.*, 2013] Liang Du, Yi-Dong Shen, Zhiyong Shen, Jianying Wang, and Zhiwu Xu. A self-supervised framework for clustering ensemble. In *Web-Age Information Management*, pages 253–264. Springer, 2013.

[Fern and Brodley, 2004] Xiaoli Zhang Fern and Carla E Brodley. Solving cluster ensemble problems by bipartite graph partitioning. In *Proceedings of the twenty-first international conference on Machine learning*, page 36. ACM, 2004.

[Grave *et al.*, 2011] Edouard Grave, Guillaume R. Obozinski, and Francis R. Bach. Trace lasso: a trace norm regularization for correlated designs. In J. Shawe-taylor, R.s. Zemel, P. Bartlett, F.c.n. Pereira, and K.q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2187–2195, 2011.

[He *et al.*, 2014] Ran He, Tieniu Tan, and Liang Wang. Robust recovery of corrupted low-rankmatrix by implicit regularizers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(4):770–783, 2014.

[Lee and Seung, 2000] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *In NIPS*, pages 556–562. MIT Press, 2000.

[Li and Ding, 2008] Tao Li and Chris Ding. Weighted consensus clustering. *Mij*, 1(2), 2008.

[Li *et al.*, 2007] Tao Li, Chris Ding, and Michael I Jordan. Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 577–582. IEEE, 2007.

[Liu *et al.*, 2010] Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In *Proceedings of the 28th International Conference on Machine Learning (ICML-10)*, 2010.

[Liu *et al.*, 2013] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):171–184, 2013.

[Luo *et al.*, 2012] Dijun Luo, Heng Huang, Feiping Nie, and Chris H Ding. Forging the graphs: A low rank and positive semidefinite graph learning approach. In *Advances in Neural Information Processing Systems*, pages 2960–2968, 2012.

[Ng *et al.*, 2001] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pages 849–856, 2001.

[Strehl and Ghosh, 2003] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3:583–617, 2003.

[Topchy *et al.*, 2003] Alexander Topchy, Anil K Jain, and William Punch. Combining multiple weak clusterings. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 331–338. IEEE, 2003.

[Topchy *et al.*, 2004] Alexander P Topchy, Anil K Jain, and William F Punch. A mixture model for clustering ensembles. In *SDM*, pages 379–390. SIAM, 2004.

[Wang *et al.*, 2009] Fei Wang, Xin Wang, and Tao Li. Generalized cluster aggregation. In *IJCAI*, pages 1279–1284, 2009.

[Wang *et al.*, 2011] Hongjun Wang, Hanhuai Shan, and Arindam Banerjee. Bayesian cluster ensembles. *Statistical Analysis and Data Mining*, 4(1):54–70, 2011.

[Xia *et al.*, 2014] Rongkai Xia, Yan Pan, Lei Du, and Jian Yin. Robust multi-view spectral clustering via low-rank and sparse decomposition. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.