Improved MinMax Cut Graph Clustering with Nonnegative Relaxation

Feiping Nie, Chris Ding, Dijun Luo, and Heng Huang

Department of Computer Science and Engineering,
University of Texas, Arlington, America
{feipingnie,dijun.luo}@gmail.com,{chqding,heng}@uta.edu

Abstract. In graph clustering methods, MinMax Cut tends to provide more balanced clusters as compared to Ratio Cut and Normalized Cut. The traditional approach used spectral relaxation to solve the graph cut problem. The main disadvantage of this approach is that the obtained spectral solution has mixed signs, which could severely deviate from the true solution and have to resort to other clustering methods, such as K-means, to obtain final clusters. In this paper, we propose to apply additional nonnegative constraint into MinMax Cut graph clustering and introduce novel algorithms to optimize the new objective. With the explicit nonnegative constraint, our solutions are very close to the ideal class indicator matrix and can directly assign clusters to data points. We present efficient algorithms to solve the new problem with the nonnegative constraint rigorously. Experimental results show that our new algorithm always converges and significantly outperforms the traditional spectral relaxation approach on ratio cut and normalized cut.

Keywords: Spectral clustering, Normalized cut, MinMax cut, Nonnegative relaxation, cluster balance, random graphs.

1 Introduction

Clustering is an important task in machine learning and data mining areas. In the past decades, many clustering algorithms have been proposed such as K-means clustering, spectral clustering and its variants [1,2,3], support vector clustering [4], and maximum margin clustering [5,6,7]. Among them, the use of manifold information in graph cut clustering has shown the state-of-the-art clustering performance and been widely applied into many applications, such as image segmentation [8], white matter fiber tracking in biomedical image [9], and protein sequence clustering [10].

MinMax Cut was proposed in [11] and showed more compact and balanced clustering results than Ratio Cut [12] and Normalized Cut [8]. Because, in MinMax Cut method, the within-cluster similarities are explicitly maximized. Solving the graph cut clustering problem is a nontrivial task. The main difficulty of the graph clustering problem lies in the constraints on the solution. In order to make the problem tractable, the constraints should be relaxed. Traditional

J.L. Balcázar et al. (Eds.): ECML PKDD 2010, Part II, LNAI 6322, pp. 451–466, 2010.

approach used spectral relaxation to solve this problem. But the main disadvantage of this approach is that the obtained spectral solution has mixed signs, which could severely deviate from the true solution and have to resort to other clustering methods, such as K-means, to obtain final cluster results.

In order to solve this notorious problem, in this paper, we propose a new method to optimize the MinMax Cut graph clustering with additional nonnegative constraint. With the explicit nonnegative constraint, the solutions are very close to the ideal class indicator matrix and can be directly used to assign cluster labels to data points. We propose efficient algorithms to solve this problem with the nonnegative constraint rigorously. Experimental results show that our algorithm always converges and the performance is significantly improved in comparisons with the traditional spectral relaxation approach on Ratio Cut and Normalized Cut.

The rest of this paper is organized as follows. Section 2 reviews the MinMax Cut problem. Our proposed nonnegative relaxation approaches to solve the MinMax Cut clustering problem are introduced in Section 3. Experimental results on real-world data sets are reported in Section 4. Finally, we conclude our work in Section 5.

2 MinMax Cut for Clustering

Suppose we have n data points $\{x_1, x_2, \cdots, x_n\}$, and construct a graph using the data with weight matrix $W \in \mathbb{R}^{n\times n}$. The multi-way MinMax Cut graph clustering objective function is (we also show Min Cut, Ratio Cut and Normalized Cut for comparisons):

$$J = \sum_{1\le p<q\le K} \frac{s(C_p, C_q)}{\rho(C_p)} + \frac{s(C_p, C_q)}{\rho(C_q)} = \sum_{k=1}^{K} \frac{s(C_k, \bar{C}_k)}{\rho(C_k)}, \tag{1}$$

$$\rho(C_k) = \begin{cases} 1 & \text{for Min Cut} \\ |C_k| & \text{for Ratio Cut} \\ \sum_{i\in C_k} d_i & \text{for Normalized Cut} \\ s(C_k, \bar{C}_k) & \text{for MinMax Cut} \end{cases} \tag{2}$$

where K is the number of clusters, C_k is the k-th cluster (subgraph in graph G), $\bar{C}_k$ is the complement of subset C_k in graph G, and $s(A, B) = \sum_{i\in A}\sum_{j\in B} W_{ij}$, $d_i = \sum_j W_{ij}$.

Let $\mathbf{q}_k (k = 1, 2, \cdots, K)$ be the cluster indicators where the i-th element of $\mathbf{q}_k$ is 1 if the i-th data point x_i belongs to cluster k, and 0 otherwise. For example, if data points within each cluster are adjacent, then

$$\mathbf{q}_k = (0, \cdots, 0, \overbrace{1, \cdots, 1}^{n_k}, 0, \cdots, 0)^T. \tag{3}$$

We can easily see that $s(C_k, \bar{C}_k) = \sum_{i\in C_k}\sum_{j\in\bar{C}_k} W_{ij} = \mathbf{q}_k^T(D-W)\mathbf{q}_k$, $\sum_{i\in C_k} d_i = \mathbf{q}_k^T D\mathbf{q}_k$, $s(C_k, C_k) = \mathbf{q}_k^T W\mathbf{q}_k$, where D is a diagonal matrix with the i-th diagonal element as d_i. We rewrite the objective functions of these four methods as:

$$J_{\text{mincut}} = \sum_{k=1}^{K} \mathbf{q}_k^T(D-W)\mathbf{q}_k, \quad J_{\text{rcut}} = \sum_{k=1}^{K} \frac{\mathbf{q}_k^T(D-W)\mathbf{q}_k}{\mathbf{q}_k^T\mathbf{q}_k}, \tag{4}$$

$$J_{\text{ncut}} = \sum_{k=1}^{K} \frac{\mathbf{q}_k^T(D-W)\mathbf{q}_k}{\mathbf{q}_k^T D\mathbf{q}_k}, \quad J_{\text{MMC}} = \sum_{k=1}^{K} \frac{\mathbf{q}_k^T(D-W)\mathbf{q}_k}{\mathbf{q}_k^T W\mathbf{q}_k}. \tag{5}$$

2.1 Cluster Balance Analysis on Random Graphs

One important advantage of the `MinMaxCut` method is that it tends to produce balanced clusters in graph clustering, *i.e.*, the resulting subgraphs will have similar size. Here we study the clustering solutions on two popular random graphs: (1) Erdos-Renyi (ER) random graph model [13,14] and (2) Expected degree sequence (EDS) random graph [15].

Erdos-Renyi Random Graph. The ER random graph model is perhaps the mostly wide used random graph model. This is a uniformly distributed random graph with n nodes, where two nodes are connected with probability $p, 0 \le p \le 1$. Considering the four objective functions, `MINcut`, `Rcut`, `Ncut`, and `MinMaxCut`, we have the following result.

Theorem 1. *For random graphs,* `MinCut` *favors highly skewed cuts.* `MinMaxCut` *favors balanced cut, i.e., both subgraphs have the same size.* `RatioCut` *and* `NormCut` *show no size preferences, i.e., each subgraph could have arbitrary size.*

Proof. We compute the object functions for the partition of G into A and B. Note that the number of edges between A and B are $p|A||B|$ on average. For `MINcut`, we have $J_{\text{mincut}}(A,B) = p|A||B|$. Clearly, `MinCut` favors either $|A| = n-1$ and $|B| = 1$, or $|B| = n-1$ and $|A| = 1$; both are skewed cuts. For `MinMaxCut`, we have

$$J_{\text{MMC}}(A,B) = \frac{|B|}{|A|-1} + \frac{|A|}{|B|-1}.$$

Minimizing $J_{\text{MMC}}(A,B)$, we obtain a balanced cut: $|A| = |B| = n/2$. For `Rcut`, we have

$$J_{\text{rcut}}(A,B) = \frac{p|A||B|}{|A|} + \frac{p|A||B|}{|B|} = np.$$

For `Ncut`, because all nodes have the same degree $(n-1)p$,

$$J_{\text{ncut}}(A,B) = \frac{p|A||B|}{p|A|(n-1)} + \frac{p|A||B|}{p|B|(n-1)} = \frac{n}{n-1}.$$

Both Rcut and Ncut objectives have no size dependency and no size preference. This random graph model shows that `MinMaxCut` has the tendency of produce a balanced clustering. □

Expected Degree Sequence Random Graph
The degree of a node on a graph is the sum of edge connecting to it. The distribution of the n node degrees of a graph is a critical property. The ER random graph has a degree distribution much like a Gaussian center around the average degree $\bar{d} = np$.

However, much of biological, social and information networks/graphs have a power-law degree distribution The expected degree sequence (EDS) random graph is a graph model for these networks/graphs. In this model, the degrees of each nodes $(d_1 \cdots d_n)$ are pre-specified. The edges are are then randomly distributed, with the constraints that the degrees of nodes satisfy the given fixed degree sequence.

The EDS graph model is a generalization of the ER random graph model, which can be seen as an special case of the EDS random graph by setting $d_i = np$ for all nodes.

For the EDS random graph model,

$$P(W_{ij} = 1) = d_i d_j / M,\ M = \sum_{ij} W_{ij}.$$

[in contrast for ER model, $P(W_{ij} = 1) = p$]. Here, to make things precise, we study a graph whose edge weights are the **average** of the probabilistic distribution: to make things precise:

$$\widehat{W}_{ij} = 1 * P(W_{ij} = 1) + 0 * P(W_{ij} = 0) = P(W_{ij} = 1) = d_i d_j / M, \qquad (6)$$

For notational simplicity, we replace $\widehat{W}$ by W. We have the following

Theorem 2. *For a EDS random graph, Rcut produces highly skewed cuts, MinMaxCutfavors balanced cut, while Ncut has no unique solution and thus shows no size preferences.*

Proof. Suppose we cut the graph into A, B. Then $S(A, B) = \sum_{i \in A} \sum_{j \in B} W_{ij}/M = \sum_{i \in A} \sum_{j \in B} d_i d_j / M = D(A)D(B)/M$. Thus for Ncut,

$$J_{\text{ncut}} = S(A, B)/D(A) + S(A, B)/D(B) + = D(B)/M + D(A)/M = 1,$$

showing no dependence on A, B. Thus Ncut has no unique solution and shows no size preference.

For Rcut and MinMaxCut, we sort the degrees in increasing order and assuming $d_1 < d_2 < d_3 < \cdots < d_n$ (we assume the degrees are different for simplicity). It is easy to see that if $|A| = k, |B| = n - k$ for fixed k, the cut $S(A, B) = D(A)D(B)/M$ is minimized when the graph $G = (v_1 \cdots v_n)$ is cut into $A = \{v_1, \cdots, v_k\}$, $B = \{v_{k+1}, \cdots, v_n\}$. Thus the optimal clustering solution is obtained by searching the minima in the range $k = 1, \cdots, n-1$. The Rcut objective is

$$J_{\text{rcut}}(k) = \frac{1}{M^2} \left[\sum_{i=1}^{k} d_i \right] \left[\sum_{r=k+1}^{n} d_r \right] \left[\frac{1}{k} + \frac{1}{n-k} \right].$$

The clustering solution is found as we search the minima in $k = 1, \cdots, n-1$. Normally, the optimal k^* is very small, $k^* \ll n$, implying a skewed cut. For MinMaxCut, $S(A,A) = \sum_{i\in A}\sum_{j\in A} W_{ij}/M = D(A)^2/M$ and $S(B,B) = D(B)^2/M$. The MinMaxCutobjective becomes

$$J_{\text{MMC}} = \frac{D(A)D(B)}{D(A)^2} + \frac{D(A)D(B)}{D(B)^2} = \frac{D(B)}{D(A)} + \frac{D(A)}{D(B)}.$$

This is minimized when $D(A) = D(B)$. The solution is generally balanced. □

Theorems 1 and 2 show the general tendency regarding to cluster balance for Rcut, Ncut and MinMaxCut. on random graphs. For pure random graphs, there is no true clusters and the clusters obtained are not meaningful. In real applications, graphs are not random. But the general tendency regarding to cluster balance are expected to be similar to Theorems 9 and 10. The following graph examples illustrate these tendencies.

We give two examples to illustrate these tendency in Figure 1, from which we can clearly see that MinMaxCutfavors more balanced cut than Rcut and Ncut.

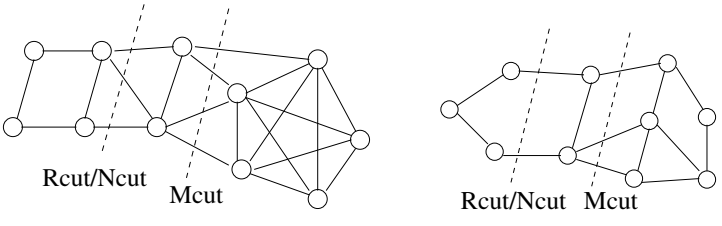


Fig. 1. Two examples that Rcut and Ncut lead to unbalanced clusters, whereas MinMaxCutgives out balanced clusters

2.2 Optimization of MinMax Cut with Spectral Relaxation

We rewrite the MinMaxCutclustering objective J_{MMC} of Eq. (5) by defining

$$Z = (\mathbf{z}_1, \cdots, \mathbf{z}_K),\ \mathbf{z}_k = \frac{\mathbf{q}_k}{\|D^{1/2}\mathbf{q}_k\|}, \tag{7}$$

the MinMaxCutclustering optimization becomes

$$\min_Z J_{\text{MMC}} = \sum_{\ell=1}^{K} \frac{1}{\mathbf{z}_\ell W \mathbf{z}_\ell} - K,\ s.t.\ Z^T DZ = I,\ Z \geq 0. \tag{8}$$

Ignoring the nonnegative constraints $Z \geq 0$ here, we derive the spectral solution.

Using Lagrangian multiplier $\Gamma = \Gamma^T$ to enforce $Z^T DZ = I$, we minimize $\mathcal{L} = J_{\text{MMC}} + \text{Tr}\ \Gamma(Z^T DZ - I)$. Setting $\partial\mathcal{L}/\partial\mathbf{z}_k = 0$, we obtain

$$\frac{W\mathbf{z}_k}{(\mathbf{z}_k^T W \mathbf{z}_k)^2} = \sum_{l=1}^{K} D\mathbf{z}_l \Gamma_{lk} \tag{9}$$

Multiply $\mathbf{z}_l^T$ from the left, we obtain

$$\Gamma_{lk} = \mathbf{z}_l^T W \mathbf{z}_k / (\mathbf{z}_k^T W \mathbf{z}_k)^2, \tag{10}$$

which is not symmetric: $\Gamma_{lk} \neq \Gamma_{kl}$. By definition Γ_{lk} must be symmetric. This implies either (1) Γ is diagonal: $\Gamma_{lk} = \delta_{kl}\Gamma_{kk}$ or (2) $\mathbf{z}_k^T W \mathbf{z}_k = \mathbf{z}_l^T W \mathbf{z}_l, k \neq l$. Condition (2) would render the objective function J_{MMC} of Eq. (8) a constant and thus is impossible. Thus we are left with the only possibility that Γ is diagonal, which in turn implies

$$\mathbf{z}_l^T W \mathbf{z}_k = \delta_{lk}\gamma_k, \tag{11}$$

and $\Gamma_{lk} = \delta_{kl}\Gamma_{kk} = \frac{\delta_{kl}\gamma_k}{\gamma_k^2} = \delta_{kl}\gamma_k^{-1}$. Now with Lagrangian multiplier Γ solved, Eq. (9) becomes

$$W\mathbf{z}_k = \gamma_k^{-1} D\mathbf{z}_k. \tag{12}$$

which is identical to the generalized eigensystem of

$$(D - W)\mathbf{z}_k = \zeta_k D\mathbf{z}_k, \tag{13}$$

where $\zeta_k = 1 - 1/\gamma_k$. Thus the solutions are given by the eigenvectors $(\mathbf{z}_1, \cdots, \mathbf{z}_K)$ of generalized Laplacian as same as Normalized Cut. Since the solution is an relaxed solution of the original minimization problem of Eq. (5), *i.e.*, an optimal solution with enlarged domain from vigorous cluster indicators Q to continuous mixed sign Z, the obtained optimal objective function value must be a lower bound for the true MinMax Cut objective

$$\sum_{k=1}^{K} \frac{1}{1 - \zeta_k} - K \leq J_{\text{MMC}}. \tag{14}$$

3 Optimization of MinMax Cut with Nonnegative Relaxation

To solve MinMax Cut problem, the traditional spectral relaxation approach relaxes the solution from binary value to real value. However, this relaxation could make the solution severely deviate from the true solution. Moreover, under this relaxation, the obtained spectral solution cannot be directly used to assign cluster labels for data points. To perform clustering, a commonly used post-processing method is to apply K-means to the space of the spectral solution to obtain clusters.

In this section, we will explicitly constrain the solution $\mathbf{q}_k$ to be nonnegative, and propose efficient algorithms to optimize the MinMax Cut clustering objective function with the nonnegative constraint on $\mathbf{q}_k$ rigorously.

3.1 Orthonormal and Nonnegative Constraints

The main difficulty of the graph clustering problem lies in the constraints of the class indicator matrix Q. The constraints should be relaxed to make the problem

solvable. From the definition of the class indicator matrix we can see that only one element is one and others are zeros in each row of Q. Thus the columns of Q are orthogonal and the orthogonality should be preserved in a relaxation of class indicator matrix. Note that the objective of the graph cuts is invariant to the scale of the columns of Q, so traditional spectral relaxation approach relaxes the constraints of Q to the orthonormal constraints:

$$Q^T Q = I. \tag{15}$$

Such relaxation makes the problem to be easily solved, but the obtained solution has mixed signs, which deviates from the class indicator matrix largely.

Note that the class indicator matrix is a nonnegative matrix, a more accurate relaxation is adding the nonnegative constraints on the Q:

$$Q^T Q = I,\ Q \geq 0. \tag{16}$$

One can see that when orthonormal and nonnegative constraints are satisfied simultaneously, only one element is positive and others are zeros in each row of Q, which is very close to the ideal class indicator matrix, and can be used directly to assign cluster labels to data points. This motivates our nonnegative relaxation approach for the MinMax Cut clustering problem, and to solve the following optimization problem:

$$\min_{Q} \sum_{k=1}^{K} \frac{\mathbf{q}_k^T D \mathbf{q}_k}{\mathbf{q}_k^T W \mathbf{q}_k},\ \ s.t.\ \ Q^T Q = I,\ Q \geq 0. \tag{17}$$

We are going to introduce two efficient algorithms to solve this problem in next subsections. The first algorithm iteratively optimizes the objective function with good performance, and the second one is more concise and also has comparable clustering results.

3.2 An Iterative Algorithm to Solve the Problem with Orthonormal and Nonnegative Constraints

In some cases, minimizing an objective might result in numerical instability [16]. Thus we turn to solve the following identical problem:

$$\max_{Q} J(Q),\ \ s.t.\ \ Q^T Q = I,\ Q \geq 0, \tag{18}$$

where

$$J(Q) = \rho \mathrm{Tr}\ Q^T Q - \sum_{k=1}^{K} \frac{\mathbf{q}_k^T D \mathbf{q}_k}{\mathbf{q}_k^T W \mathbf{q}_k}, \tag{19}$$

ρ is an appropriate positive value such that $\rho W - (D - W)$ is positive semi-definite. Specifically, we set $\rho = \frac{\lambda_{max}(D-W)}{\lambda_{max}(W)}$ in this work, where $\lambda_{max}(W)$ and $\lambda_{max}(D - W)$ denotes the largest eigenvalue of W and $D - W$, respectively.

We begin with the Lagrangian function

$$\mathcal{L} = J(Q) - \mathrm{Tr}\Lambda(Q^TQ - I) - \mathrm{Tr}\Sigma^TQ, \tag{20}$$

where the Lagrange multiplier Λ enforces the orthogonality condition $Q^TQ = I$ and the Lagrange multiplier Σ enforces the nonnegativity condition $Q \geq 0$. Using the KKT complementary slackness condition, we have

$$(\frac{\partial J}{\partial Q} - 2Q\Lambda)_{ik}Q_{ik} = 0. \tag{21}$$

Summing over k, we obtain $(\frac{1}{2}Q^T\frac{\partial J}{\partial Q})_{ii} = (Q^TQ\Lambda)_{ii} = \Lambda_{ii}$. This gives the diagonal elements of Λ. To find the off-diagonal elements of Λ, we temporarily ignore the nonnegativity condition, which gives $(\frac{\partial J}{\partial Q} - 2Q\Lambda)_{ik} = 0$. Left multiplying by $Q_{i'k}$ and summing over k, we obtain $(\frac{1}{2}Q^T\frac{\partial J}{\partial Q})_{i'i} = \Lambda_{i'i}$ for the off-diagonal elements of Λ. Combining these two results yields

$$\Lambda = \frac{1}{2}Q^T\frac{\partial J}{\partial Q}. \tag{22}$$

Note that

$$\frac{1}{2}\frac{\partial J}{\partial Q} = \rho Q - DQ_\alpha + WQ_\beta, \tag{23}$$

where

$$Q_\alpha = \left[\frac{1}{\mathbf{q}_1^TW\mathbf{q}_1}\mathbf{q}_1, \cdots, \frac{1}{\mathbf{q}_K^TW\mathbf{q}_K}\mathbf{q}_K\right], \tag{24}$$

$$Q_\beta = \left[\frac{(\mathbf{q}_1^TD\mathbf{q}_1)}{(\mathbf{q}_1^TW\mathbf{q}_1)^2}\mathbf{q}_1, \cdots, \frac{(\mathbf{q}_K^TD\mathbf{q}_K)}{(\mathbf{q}_K^TW\mathbf{q}_K)^2}\mathbf{q}_K\right], \tag{25}$$

then we have

$$\Lambda = \rho Q^TQ - Q^TDQ_\alpha + Q^TWQ_\beta. \tag{26}$$

Decomposing Λ into positive part and negative part as

$$\Lambda = \Lambda^+ - \Lambda^-, \tag{27}$$

where $\Lambda^+ = (|\Lambda|+\Lambda)/2$ and $\Lambda^- = (|\Lambda|-\Lambda)/2$. Now concentrating on the variable Q while ignoring constant terms in $\mathcal{L}$, we have

$$\begin{aligned}\frac{1}{2}\frac{\partial(J - \mathrm{Tr}\Lambda Q^TQ)}{\partial Q} &= \rho Q - DQ_\alpha + WQ_\beta - Q\Lambda \\ &= \rho Q - DQ_\alpha + WQ_\beta - Q\Lambda^+ + Q\Lambda^- \\ &= (\rho Q + WQ_\beta + Q\Lambda^-) - (DQ_\alpha + Q\Lambda^+).\end{aligned} \tag{28}$$

As in Nonnegative Matrix Factorization (NMF) [17,18], Eq. (28) leads to the following multiplicative update formula:

$$Q_{ik} \leftarrow Q_{ik}\sqrt{\frac{(\rho Q + WQ_\beta + Q\Lambda^-)_{ik}}{(DQ_\alpha + Q\Lambda^+)_{ik}}}. \tag{29}$$

We can see that using this update, Q_{ik} will increase when the corresponding element of the gradient in Eq. (28) is larger than zero, and will decrease otherwise. Therefore, the update direction is consistent to the update direction in the gradient ascent method. Our extensive experiments show that the iterative algorithm presented here always converges and monotonically increases the objective in each iteration. The computational cost in each iteration is of $O(n^2)$, which is similar to traditional spectral clustering algorithm.

As mentioned before, the solution is very close to the ideal class indicator matrix due to the orthonormal and nonnegative constraints. Thus the solution Q can be directly used to assign cluster labels to data points. Specifically, the i-th data point x_i is assigned to cluster label c_i as $c_i = \arg\max_k Q_{ik}$.

3.3 Initialization for the Iterative Algorithm

From the update formula in Eq. (29), we can see that if the initialization of Q is nonnegative, then Q will preserve nonnegative in the update process, and hence the nonnegative constraint of the solution is naturally satisfied. As the spectral relaxation of MinMax Cut problem is identical to the spectral relaxation of Normalized Cut problem, we initialize Q by $Q_0 + 0.2$, where Q_0 is obtained by spectral relaxation of Normalized Cut followed by K-means clustering in the eigenspace. Note that Q_0 is a cluster indicator matrix, and the initialization should not be a cluster indicator matrix (otherwise the values won't change during the iteration), thus we plus 0.2 in practice. It is worth noting that the initialization is not very sensitive, we can also use random initialization as well but the result would be more stable if using the initialization suggested here.

3.4 A New Concise Algorithm

In this section, we propose a more concise NMF algorithm to solve the MinMaxCutproblem. We start with the Eqs. (7, 8) formulation. Using Lagrangian multiplier $\Omega = \Omega^T$ to enforce $Z^T D Z = I$, we minimize

$$\mathcal{L}(Z) = J_{\text{MMC}} + \text{Tr } \Omega(Z^T D Z - I). \tag{30}$$

The KKT complementary slackness condition for the nonnegativity condition $Z \geq 0$ gives [noting $Z_{ik} = (\mathbf{z}_k)_i$]

$$0 = \frac{\partial \mathcal{L}}{\partial Z_{ik}} Z_{ik} = \left(-\frac{(W\mathbf{z}_k)_i}{(\mathbf{z}_k^T W \mathbf{z}_k)^2} + (DZ\Omega)_{ik} \right) Z_{ik}. \tag{31}$$

Summing over i, we obtain

$$\Omega_{kk} = \frac{1}{\mathbf{z}_k^T W \mathbf{z}_k} = \frac{1}{(Z^T W Z)_{kk}} \tag{32}$$

To find the off-diagonal elements of Ω, we look at the Lagrangian multiplier for the case where the nonnegativity constraint is ignored, which is given in Eq. (10).

From Eq. (10). we propose three strategies to obtain a symmetrized Ω as follows:

$$S_1: \quad \Omega_{lk} = \frac{(Z^T W Z)_{lk}}{(Z^T W Z)_{kk}(Z^T W Z)_{ll}} \tag{33}$$

$$S_2: \quad \Omega_{lk} = \frac{(Z^T W Z)_{lk}}{2(Z^T W Z)^2_{kk}} + \frac{(Z^T W Z)_{lk}}{2(Z^T W Z)^2_{ll}} \tag{34}$$

$$S_3: \quad \Omega_{lk} = \frac{(Z^T W Z)_{lk}}{\frac{1}{2}(Z^T W Z)^2_{kk} + \frac{1}{2}(Z^T W Z)^2_{ll}}. \tag{35}$$

Note that all these formulas reduce to Eq. (32) when $l = k$. The gradient decent algorithm is

$$Z_{ik} \leftarrow Z_{ik} - \eta_{ik}\frac{\partial \mathcal{L}}{\partial Z_{ik}} = Z_{ik} - \eta_{ik}\left(-\frac{(WZ)_{ik}}{(Z^T W Z)^2_{kk}} + (DZ\Omega)_{ik}\right) \tag{36}$$

Setting $\eta_{ik} = Z_{ik}/(DZ\Omega)_{ik}$ leads to the update formula

$$Z_{ik} \leftarrow Z_{ik}\sqrt{\frac{(WZ)_{ik}}{(DZ\Omega)_{ik}(Z^T W Z)^2_{kk}}}. \tag{37}$$

Our extensive experiment results show that the iterative algorithm using any one of three above symmetrization strategies always converges and monotonically decreases the objective $\mathcal{L}(Z)$ in each iteration.

4 Experimental Results

In this section, we will evaluate the effectiveness of the proposed nonnegative relaxation algorithms for MinMax Cut graph clustering on eight benchmark data sets. We also compare the clustering performance of our algorithms to the traditional spectral relaxation algorithm for Ratio Cut [12] and for Normalized Cut [8] graph clustering, respectively.

4.1 Experimental Setup

Eight benchmark data sets are used in our experiments, including two UCI data sets [1] (Ecoli and Vehicle), one character data set, Binalpha, one object data set, COIL-20 [19], and four face image data sets, Yale, AT&T [20], Umist [21], and YaleB [22]. Some data sets are resized, and Table 1 summarizes the details of all data sets used in the experiments.

We use Gaussian function to construct the weight matrix W. The weight W_{ij} is defined as

$$W_{ij} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right) & x_i \text{ and } x_j \text{ are neighbors;} \\ 0 & \text{otherwise.} \end{cases} \tag{38}$$

[1] http://www.ics.uci.edu/∼mlearn/MLRepository.html

Table 1. Data set Descriptions

Data set	Size	Dimensions	Classes	Data set	Size	Dimensions	Classes
Ecoli	336	343	8	AT&T	400	644	40
Vehicle	846	18	4	Umist	575	644	20
Binalpha	1404	320	36	Coil20	1440	1024	20
Yale	165	3456	15	YaleB	2414	1024	38

Table 2. The cluster balance and clustering accuracy of Ratio Cut, Normalized Cut, and the proposed Nonnegative MinMax Cut in Section 3.2. The values after '±' are the standard deviations.

Data set	Ratio Cut		Normalized Cut		MinMax Cut	
	Balance	Accuracy	Balance	Accuracy	Balance	Accuracy
Ecoli	8.41	57.59 ± 2.31 %	6.17	55.95 ± 1.75 %	**5.31**	**58.30 ± 0.85** %
Vehicle	49.51	40.66 ± 0.57 %	56.14	43.74 ± 1.22 %	**49.13**	**44.40 ± 0.91** %
Binalpha	16.29	44.85 ± 1.96 %	24.56	44.29 ± 1.26 %	**9.26**	**46.32 ± 1.32** %
Yale	13.25	65.39 ± 3.96 %	1.57	70.06 ± 0.31 %	**1.13**	**71.52 ± 0.00** %
AT&T	11.15	70.75 ± 2.19 %	3.69	75.92 ± 1.17 %	**2.14**	**79.88 ± 1.13** %
Umist	5.82	60.00 ± 2.86 %	5.78	60.59 ± 1.14 %	**4.17**	**62.92 ± 0.93** %
Coil20	11.35	71.13 ± 4.88 %	6.11	78.19 ± 1.76 %	**5.02**	**79.09 ± 2.18** %
YaleB	49.18	38.55 ± 0.98 %	68.38	39.66 ± 1.35 %	**41.71**	**45.08 ± 1.32** %

The number of neighbors and the parameter σ should be predefined by user. In our experiments, we set the number of neighbors to be 5 (that is a commonly used number) in all data sets, and use the self-tune spectral clustering [23] method to determine the parameter σ.

4.2 Evaluation Metrics

We use the following two standard evaluation metrics to evaluate the performance for the three graph cut clustering algorithms.

Cluster Balance is defined as:

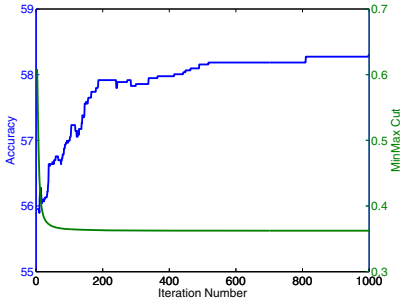
$$CB = \frac{N_{max} - N_{min}}{N_{min}},$$

where N_{max} is the number of data points in the cluster with largest size, and N_{min} is the number of data points in the cluster with smallest size. A smaller CB indicates a more balanced clustering.

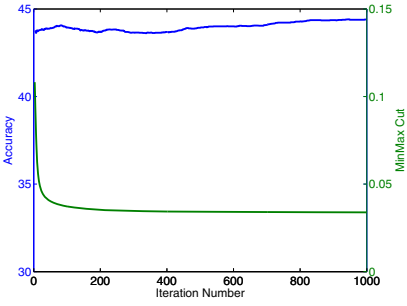
Clustering Accuracy is calculated by:

$$ACC = \frac{\sum_{i=1}^{n} \delta(l_i, map(c_i))}{n},$$

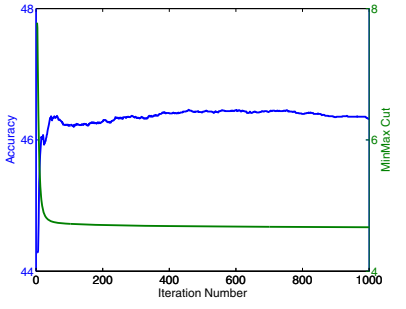
where l_i is the true class label and c_i is the obtained cluster label of x_i, $\delta(x, y)$ is the delta function, and $map(\cdot)$ is the best mapping function. Note $\delta(x, y) = 1$,



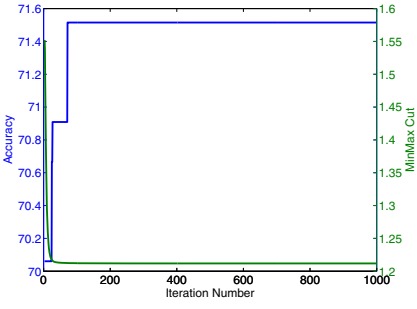
(a) Ecoli



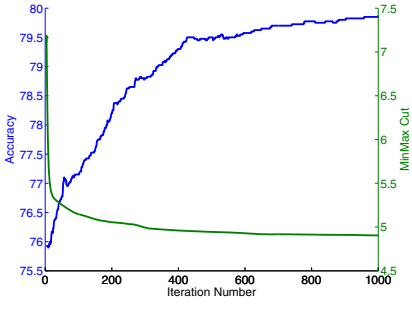
(b) Vehicle



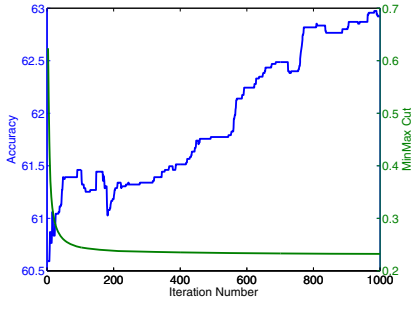
(c) Binalpha



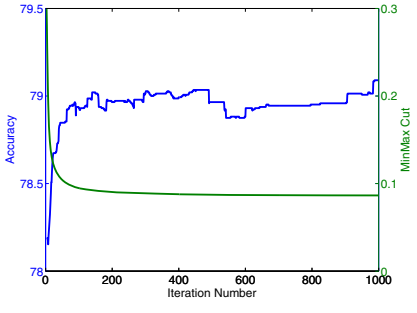
(d) Yale



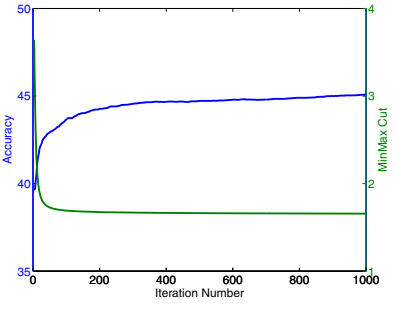
(e) AT&T



(f) Umist

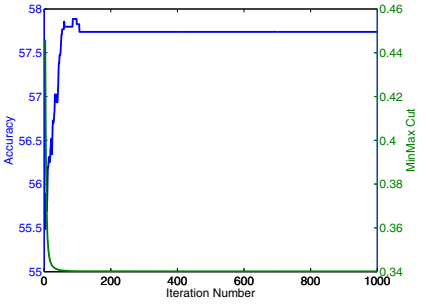


(g) Coil20

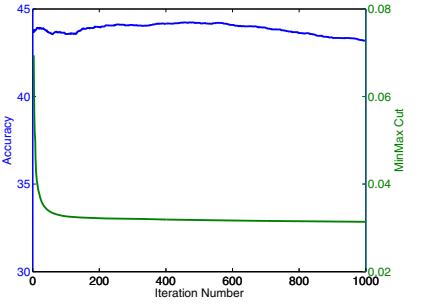


(h) YaleB

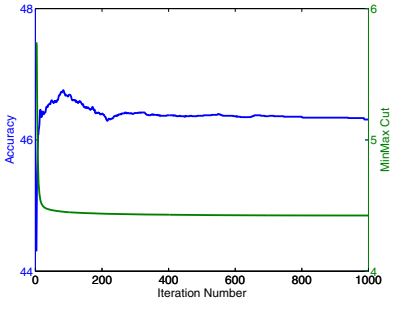
Fig. 2. The variation process of the clustering accuracy and the MinMax Cut objective value *w.r.t.* iteration number for the algorithm proposed in Section 3.2



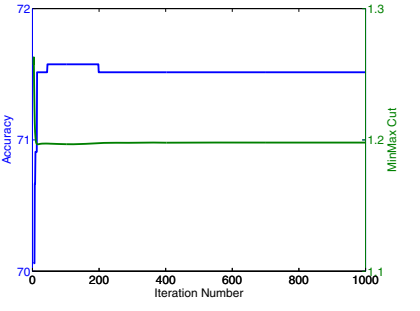
(a) Ecoli



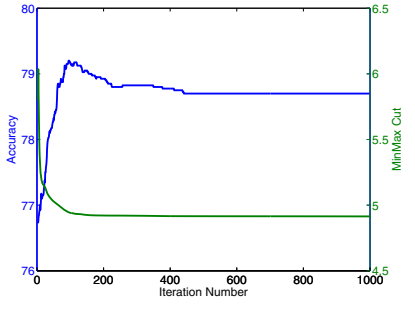
(b) Vehicle



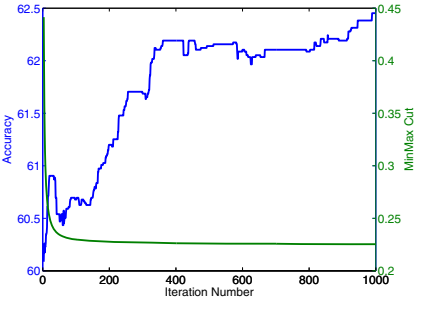
(c) Binalpha



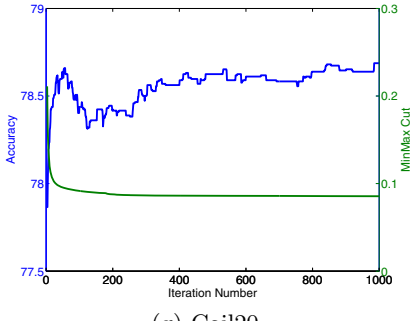
(d) Yale



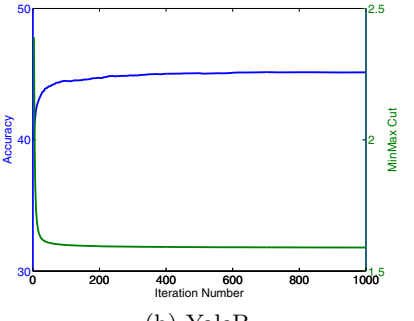
(e) AT&T



(f) Umist



(g) Coil20



(h) YaleB

Fig. 3. The variation process of the clustering accuracy and the MinMax Cut objective value *w.r.t.* iteration number for the concise algorithm proposed in Section 3.4 with the first symmetrization strategy

Table 3. Cluster balance and clustering accuracy of the concise algorithm proposed in Section 3.4 with the three symmetrization strategies

Data set	Strategy 1		Strategy 2		Strategy 3	
	Balance	Accuracy	Balance	Accuracy	Balance	Accuracy
Ecoli	5.10	57.74 ± 0.00 %	5.10	57.74 ± 0.00 %	5.10	57.74 ± 0.00 %
Vehicle	47.63	43.20 ± 1.04 %	47.63	43.20 ± 1.04 %	47.63	43.20 ± 1.04 %
Binalpha	7.83	46.31 ± 1.89 %	7.83	46.31 ± 1.89 %	7.72	46.31 ± 1.89 %
Yale	1.13	71.52 ± 0.00 %	1.13	71.52 ± 0.00 %	1.13	71.52 ± 0.00 %
ORL	2.54	78.70 ± 1.81 %	2.54	78.70 ± 1.81 %	2.54	78.70 ± 1.81 %
Umist	4.97	62.45 ± 0.52 %	4.97	62.45 ± 0.52 %	4.97	62.45 ± 0.52 %
Coil20	5.08	78.69 ± 2.21 %	5.08	78.69 ± 2.21 %	5.08	78.69 ± 2.21 %
YaleB	37.73	45.14 ± 1.21 %	37.73	45.14 ± 1.21 %	37.73	45.14 ± 1.21 %

if $x = y$; $\delta(x, y) = 0$, otherwise. The mapping function $map(\cdot)$ matches the true class label and the obtained cluster label and the best mapping is solved by Kuhn-Munkres algorithm [24]. A larger ACC indicates a better performance.

4.3 Evaluation Results

The results of all clustering algorithms depend on the initialization. To reduce statistical variation, we run the Ratio Cut algorithm and the Normalized Cut algorithm with the same 1000 random initializations. Ten results corresponding to the 10 best objective values are selected from the 1000 runs. Then we run the proposed nonnegative MinMax Cut algorithm proposed in Section 3.2 and 3.4 using the 10 results of Normalized Cut as initialization and also obtain 10 new results. We record all the ten results and the mean results are reported in the experiments. The clustering results from three graph cut methods are reported in Table 2 and 3. From the results, we have three following observations:

1) The Normalized Cut frequently, but not always, yields more balanced clustering than Ratio Cut. The Nonnegative MinMax Cut consistently yield more balanced clustering than both Normalized Cut and Ratio Cut.
2) The Normalized Cut frequently, but not always, outperforms Ratio Cut in term of clustering accuracy. The Nonnegative MinMax Cut outperforms Normalized Cut and Ratio Cut on all eight benchmark data sets, and the improvement is significant in some cases.
3) The results of algorithm proposed in Section 3.2 are slightly better than those of algorithm proposed in Section 3.4, but the latter one is simpler and does not need to calculate the ρ in Eq. (19). We can also observe that the three symmetrization strategies almost yield the same results, thus we can select anyone of them in practice.

To evaluate the convergency and effectiveness of our iterative algorithms, we plot the variation process of the clustering accuracy and the MinMax Cut objective value defined in Eq. (30) *w.r.t.* iteration number. Figure 2 shows the variation

process for the algorithm proposed in Section 3.2. From the figures, we can see that our algorithm always converges on all eight data sets, and the MinMax Cut objective value is monotonically decreased in each iteration, theoretically proving it is an interesting issue in the future work. On the other hand, the clustering accuracy tends to increase in the iteration, which indicates that the MinMax Cut objective value is consistent to the clustering accuracy, and hence is a reasonable objective for clustering problem.

Figure 3 shows the variation process for the concise algorithm proposed in Section 3.4 (as the results of the three strategies are very similar, we only show the results of the first symmetrization strategy here). From all figures, we can see that the simpler algorithms also converge on all eight data sets. The MinMax Cut objective value is monotonically decreased in each iteration, and the clustering accuracy tends to increase in each iteration.

5 Conclusions

In this paper, we proposed the nonnegative relaxation to solve the MinMax Cut graph clustering problem, and introduced efficient algorithms to solve this problem with explicit nonnegative constraint rigorously. Differing from the traditional spectral relaxation approach, the proposed nonnegative relaxation approach makes the solution very close to the ideal class indicator matrix, and can be directly used to assign cluster labels to data points. Extensive experimental results on eight benchmark data sets show that the proposed algorithms always converge and the performance is significantly improved in comparisons with the traditional spectral relaxation approach on ratio cut and normalized cut.

Acknowledgments. This research is supported by NSF-CCF 0830780, NSF-CCF 0939187, NSF-CCF 0917274, NSF-DMS 0915228, NSF-CNS 0923494.

References

1. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems (NIPS), vol. 14, pp. 849–856 (2002)
2. Nie, F., Xu, D., Tsang, I.W., Zhang, C.: Spectral embedded clustering. In: IJCAI, pp. 1181–1186 (2009)
3. Yang, Y., Xu, D., Nie, F., Yan, S., Zhuang, Y.: Image clustering using local discriminant models and global integration. IEEE Transactions on Image Process (2010)
4. Ben-Hur, A., Horn, D., Siegelmann, H., Vapnik, V.: Support vector clustering 2, 125–137 (2001)
5. Xu, L., Neufeld, J., Larson, B., Schuurmans, D.: Maximum margin clustering. MIT Press, Cambridge (2005)
6. Zhang, K., Tsang, I., Kwok, J.: Maximum margin clustering made practical. In: ICML, Corvallis, Oregon, USA (2007)

7. Li, Y., Tsang, I., Kwok, J.T., Zhou, Z.: Tighter and convex maximum margin clustering. In: AISTATS (2009)
8. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on PAMI 22(8), 888–905 (2000)
9. Brun, A., Park, H.J., Shenton, M.E.: Clustering fiber traces using normalized cuts. In: Barillot, C., Haynor, D.R., Hellier, P. (eds.) MICCAI 2004. LNCS, vol. 3216, pp. 368–375. Springer, Heidelberg (2004)
10. Pentney, W., Meila, M.: Spectral clustering of biological sequence data. In: AAAI (2005)
11. Ding, C.H.Q., He, X., Zha, H., Gu, M., Simon, H.D.: A min-max cut algorithm for graph partitioning and data clustering. In: ICDM, pp. 107–114 (2001)
12. Chan, P.K., Schlag, M.D.F., Zien, J.Y.: Spectral k-way ratio-cut partitioning and clustering. IEEE Trans. on CAD of Integrated Circuits and Systems 13(9), 1088–1096 (1994)
13. Cheng, C.K., Wei, Y.C.A.: An improved two-way partitioning algorithm with stable performance [vlsi]. IEEE Trans. on CAD of Integrated Circuits and Systems 10(12), 1502–1511 (1991)
14. Bollobas, B.: Random graphs (1985)
15. Chung, F., Lu, L.: Complex Graphs and Networks. Amer. Math. Society, Providence (2006)
16. Hou, C., Zhang, C., Wu, Y., Jiao, Y.: Stable local dimensionality reduction approaches. Pattern Recognition 42(9), 2054–2066 (2009)
17. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: NIPS, pp. 556–562. MIT Press, Cambridge (2001)
18. Li, T., Ding, C.H.Q.: The relationships among various nonnegative matrix factorization methods for clustering. In: Perner, P. (ed.) ICDM 2006. LNCS (LNAI), vol. 4065, pp. 362–371. Springer, Heidelberg (2006)
19. Nene, S.A., Nayar, S.K., Murase, H.: Columbia object image library (COIL-20), Technical Report CUCS-005-96, Columbia University (1996)
20. Samaria, F.S., Harter, A.C.: Parameterisation of a stochastic model for human face identification. In: 2nd IEEE Workshop on Applications of Computer Vision, pp. 138–142 (1994)
21. Graham, D.B., Allinson, N.M.: Characterizing virtual eigensignatures for general purpose face recognition. in face recognition: From theory to applications. NATO ASI Series F, Computer and Systems Sciences 163, 446–456 (1998)
22. Georghiades, A., Belhumeur, P., Kriegman, D.: From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Transactions on PAMI 23(6), 643–660 (2001)
23. Zelnik-Manor, L., Perona, P.: Self-tuning spectral clustering. In: NIPS (2004)
24. Lovász, L., Plummer, M.: Matching theory. Akadémiai Kiadó, Budapest (1986)