Gao, Junning; Yamada, Makoto; Kaski, Samuel; Mamitsuka, Hiroshi; Zhu, Shanfeng

A Robust Convex Formulation for Ensemble Clustering

Published in:
Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)

Published: 01/07/2016

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Gao, J., Yamada, M., Kaski, S., Mamitsuka, H., & Zhu, S. (2016). A Robust Convex Formulation for Ensemble Clustering. In S. Kambhampati (Ed.), *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)* (pp. 1476-1482). AAAI Press. http://www.ijcai.org/Proceedings/16/Papers/212.pdf

A Robust Convex Formulation for Ensemble Clustering

Junning Gao,[1] Makoto Yamada,[2] Samuel Kaski,[2,3] Hiroshi Mamitsuka,[2,3] Shanfeng Zhu[1]
[1] School of Computer Science and Shanghai Key Lab of Intelligent Information Processing
Fudan University, Shanghai, China.
[2] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Japan.
[3] Department of Computer Science, Aalto University, Finland.
13110240002@fudan.edu.cn, myamada@kuicr.kyoto-u.ac.jp, samuel.kaski@aalto.fi
mami@kuicr.kyoto-u.ac.jp, zhusf@fudan.edu.cn

Abstract

We formulate ensemble clustering as a regularization problem over nuclear norm and cluster-wise group norm, and present an efficient optimization algorithm, which we call Robust Convex Ensemble Clustering (RCEC). A key feature of RCEC allows to remove anomalous cluster assignments obtained from component clustering methods by using the group-norm regularization. Moreover, the proposed method is convex and can find the globally optimal solution. We first showed that using synthetic data experiments, RCEC could learn stable cluster assignments from the input matrix including anomalous clusters. We then showed that RCEC outperformed state-of-the-art ensemble clustering methods by using real-world data sets.

1 Introduction

Clustering is a major technique in data science with many applications including text clustering [Huang *et al.*, 2011] and image segmentation [Shi and Malik, 2000], etc. However, no single clustering algorithm, such as k-means [Hartigan and Wong, 1979] and spectral clustering [Ng *et al.*, 2002], can capture all types of patterns and so clustering results are likely to be unstable, if the assumption behind the used model is violated [Jain, 2010].

Ensemble clustering [Strehl and Ghosh, 2003] has been proposed to improve the performance of a single clustering algorithm by integrating more than one clustering results (i.e., the partition vectors $\boldsymbol{p}$, see Figure 1). Ensemble clustering can be regarded as a meta-learning algorithm [Zhou, 2012], and there exist various types of methods including graph theory based approaches [Fern and Brodley, 2004], probabilistic models [Topchy *et al.*, 2005; Wang *et al.*, 2011], matrix factorization based models [Li *et al.*, 2007], and matrix completion [Yi *et al.*, 2012]. However, these algorithms implicitly assume that all input partitions by different clustering algorithms are reasonably good while this assumption can be easily violated if partition vectors include anomalous partitions.

For solving the problem of anomalous partitions, the weighted consensus clustering (WCC) algorithm has been proposed [Li and Ding, 2008]. WCC uses "importance" weights over partitions which are estimated from inputs by quadratic programming (QP), resulting lower weights for more anomalous partitions to be removed. Recently, an alternative approach, Instance-wise weighted NMF (Nonnegative Matrix Factorization)-based Aggregation (INA), has been proposed [Zheng *et al.*, 2015]. INA tries to find anomalous *clusters* within a partition (See Figure 1), while WCC tries to find anomalous partitions. Both approaches are state-of-the-art ensemble clustering methods that can deal with anomalous partitions/clusters. However, they are non-convex; the clustering performance heavily depends on initial parameter values. That is, for both WCC and INA, parameter values need to be carefully initialized. However, finding appropriate initial parameter values is a rather hard problem.

We propose Robust Convex Ensemble Clustering (RCEC). Specifically, we formulate the ensemble clustering problem as a nuclear norm and cluster-wise group norm regularization problem. A clear advantage of RCEC over WCC and INA is that the formulation of RCEC is convex and can find the globally optimal solution, being free from tuning initial parameter values. Moreover, the group norm regularization allows to remove anomalous clusters efficiently. Through synthetic and real-world experiments, we show that the proposed method outperformed state-of-the-art ensemble clustering methods. In summary, the contributions of this paper are as follows:

- We formulate the ensemble clustering problem as an optimization problem with nuclear norm and cluster-wise group norm regularization, and propose an efficient optimization algorithm over this problem.
- We introduce the $\ell_{2,1}$ norm regularization to detect anomalous clusters. To our knowledge, this is the first work to use the $\ell_{2,1}$ norm in ensemble problems.
- The proposed method empirically outperformed state-of-the-art ensemble clustering algorithms.

2 Problem Formulation

Let $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N] \in \mathbb{R}^{d \times N}$ be the feature matrix for N instances and $\boldsymbol{P} = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_M] \in \mathbb{R}^{N \times M}$ be the input partition matrix obtained by applying M clustering algorithms

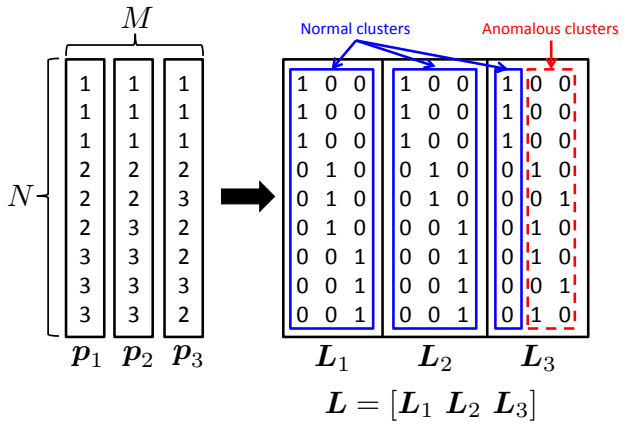


Figure 1: An example of cluster assignment matrix $\boldsymbol{L}$ ($N = 9, M = 3, K_1 = 3, K_2 = 3, K_3 = 3$).

to $\boldsymbol{S}$. The $\boldsymbol{p}_m = [p_{1,m}, p_{2,m}, \ldots, p_{N,m}]^\top \in \mathbb{R}^N$ are the partition vectors, where $p_{i,m} \in \{1, 2, \ldots, K_m\}$ and K_m is the number of clusters for the m-th partition. Note that, the number of clusters in different partitions can be different.

We further define binary cluster assignment matrices $\boldsymbol{L}_m \in \mathbb{R}^{N \times K_m}$, indicating the clusters to which each instance belongs, as follows:

$$[\boldsymbol{L}_m]_{ij} = \begin{cases} 1 & \text{if } p_{i,m} = j \\ 0 & \text{otherwise} \end{cases}.$$

The columns of the partition matrix $\boldsymbol{L}_m$ can be permuted without loss of generality.

We pool all binary cluster assignment matrices as

$$\boldsymbol{L} = [\boldsymbol{L}_1, \ldots \boldsymbol{L}_M] \in \mathbb{R}^{N \times K},\ K = \sum_{m=1}^{M} K_m$$

and use $\boldsymbol{L}$ as the input matrix for our algorithm. In this paper, we assume two types of clusters: normal clusters with dominant patterns and anomalous clusters without the patterns. Figure 1 shows an example of $\boldsymbol{L}$.

The goal of this paper is to find normal partitions/clusters from the binary cluster assignment matrices, and use the obtained normal clusters for meta-clustering.

3 Related Work

We first introduce related ensemble clustering methods and discuss their drawbacks.

Graph-based ensemble methods: Graph-based methods first construct a graph (bi-graph, hypergraph, or meta-graph) from input partitions and then apply a graph-cut partitioning algorithm to get ensemble clustering results. The hybrid bipartite graph formation (HBGF) generates a bipartite meta-graph from a binary assignment matrix $\boldsymbol{L}$, where both input clusters and instances are nodes. HBGF then finds a clustering assignment from the bipartite graph. HBGF has been empirically verified to be a state-of-the-art graph-based method [Huang *et al.*, 2011; Fern and Brodley, 2004]. However, HBGF treats normal and anomalous partitions/clusters equally.

Matrix completion based ensemble methods: ECMC [Yi *et al.*, 2012] is a two-step method. The first step is to construct the similarity matrix $\boldsymbol{W}^m$ ($\boldsymbol{W}^m_{ij} = 1$ if instances i and j are in the same cluster of the m-th partition, zero otherwise.) from a binary assignment matrix $\boldsymbol{L}$ and filter out uncertain data pairs (possibly anomalous input pairs) using predefined thresholds. Then, they use a matrix completion algorithm to complete the partially observed similarity matrix. Finally, spectral clustering is used to obtain the final clustering result from the completed similarity matrix. ECMC uses pre-defined thresholding parameters, and may lose some important information. Similarly, the Robust Clustering Ensemble (RCE) [Zhou *et al.*, 2015] also focuses on detecting anomalous instance pairs.

Non-negative matrix factorization (NMF) based ensemble methods: NMF-based methods factorize the input similarity matrix or the cluster assignment matrix, to obtain a consensus cluster assignment matrix. NMF-based consensus clustering (NMFC) [Li *et al.*, 2007] is the first NMF-based algorithm. Similar to ECMC, NMFC first computes the similarity matrix from $\boldsymbol{L}$ and then uses orthogonal non-negative matrix trifactorization (tri-NMF) [Ding *et al.*, 2006] to have a cluster assignment matrix. Since NMFC uses the input similarity matrix that is estimated from all normal and anomalous partitions, the performance of NMFC can be poor if there exist some anomalous partitions. To overcome this issue, WCC [Li and Ding, 2008] introduces "importance" weights over input partitions which are determined by quadratic programming (QP). That is, WCC computes the similarity matrix by putting larger weights for normal partitions and smaller weights for anomalous partitions. However, WCC assumes that some partition vectors $\boldsymbol{p}$ are anomalous. This means even if clusters within an anomalous partition are normal, WCC may ignore those information. Moreover, WCC is computationally expensive due to the QP computation.

Instance-wise Weighted NMF-based Aggregation (INA) [Zheng *et al.*, 2015] was proposed to overcome the problem of WCC. More specifically, INA introduces weights over clusters to explicitly consider normal/anomalous clusters. The objective function of INA is given as

$$\begin{aligned} \min_{\boldsymbol{H}\geq 0, \boldsymbol{G}\geq 0, \boldsymbol{\Psi}\geq 0} \quad & \|\boldsymbol{L} \odot \boldsymbol{\Psi} - \boldsymbol{H}\boldsymbol{G}^T\|^2 + \lambda \|\boldsymbol{\Psi}\|^2 \\ \text{s.t.} \quad & (\boldsymbol{L} \odot \boldsymbol{\Psi})\mathbf{1}^{K\times 1} = \mathbf{1}^{N\times 1}, \end{aligned}$$

where $\boldsymbol{\Psi} = (\phi_{ij})^{N\times K}$ is a weighting matrix, indicating the reliability of cluster j for instance i. WCC and INA can handle anomalous partitions/clusters, but both are non-convex, resulting in that initial parameter values need to be carefully tuned.

4 Proposed method

4.1 Robust Convex Ensemble Clustering Model

The key idea of RCEC is to reconstruct a cluster assignment matrix $\boldsymbol{X}$ from $\boldsymbol{L}$ by choosing representative *normal* clusters. To this end, we impose cluster-wise (column-wise) sparsity by assigning a $\ell_{2,1}$ regularizer to $\boldsymbol{X}$. Moreover, since the matrix $\boldsymbol{L}$ tends to be low-rank, we assume the rank of the

obtained assignment matrix to be smaller than the number of true clusters c.

The optimization problem can be written as

$$\begin{aligned} \min_{\boldsymbol{X}\geq 0} \quad & \|\boldsymbol{L}-\boldsymbol{X}\|_F^2+\beta\|\boldsymbol{X}\|_{2,1} \\ \text{s.t.} \quad & \text{rank}(\boldsymbol{X})\leq c, \end{aligned} \tag{1}$$

where $\beta \geq 0$ is the regularization parameter tuning the number of anomalous clusters, $\|\cdot\|_F$ is the Frobenius norm, and $\|\cdot\|_{2,1}$ is the $\ell_{2,1}$-norm defined as follows [Yuan and Lin, 2006; Yang *et al.*, 2011; He *et al.*, 2012; Yang *et al.*, 2013]:

$$\|\boldsymbol{X}\|_{2,1}=\sum_{j=1}^{K}\sqrt{\sum_{i=1}^{N}\boldsymbol{X}_{ij}^2}=\sum_{j=1}^{K}\|\boldsymbol{X}_{\cdot j}\|_2,$$

where $\boldsymbol{X}_{\cdot j}$ represents the j-th column of $\boldsymbol{X}$, which corresponds to a particular cluster. The $\ell_{2,1}$-norm enforces sparsity over groups and non-sparsity within a group, and thus, the coefficients of anomalous clusters can be shrunk to zero by imposing the $\ell_{2,1}$-norm regularization. Note that, we add the *non-negativity* constraint $\boldsymbol{X}\geq 0$, since the input binary partition matrix $\boldsymbol{L}$ is non-negative.

The rank constraint in Eq. (1) is non-convex, and we use the nuclear norm as a convex surrogate of the rank function. The nuclear norm is defined as

$$\|\boldsymbol{X}\|_*=\text{tr}\left((\boldsymbol{X}^\top\boldsymbol{X})^{1/2}\right)=\sum_{i=1}^{\min(N,K)}\sigma_i,$$

where σ_i is the singular values of $\boldsymbol{X}$. Since the nuclear norm is defined as the sum of singular values, it can be regarded as the ℓ_1 regularization of singular values. Thus, by imposing the nuclear norm, we can obtain low-rank estimation.

We then reformulate the objective function as the following:

$$\min_{\boldsymbol{X}\geq 0}\|\boldsymbol{L}-\boldsymbol{X}\|_F^2+\lambda\|\boldsymbol{X}\|_*+\beta\|\boldsymbol{X}\|_{2,1},$$

where $\lambda\geq 0$ is a regularization parameter.

Due to the non-differentiability of the nuclear norm, we consider the Schatten-p norm as a uniform smooth approximation to the nuclear norm [Mohan and Fazel, 2012]. The smooth Schatten-p function is given as

$$\begin{aligned} f_p(\boldsymbol{X}) &= \text{tr}\left((\boldsymbol{X}^T\boldsymbol{X}+\gamma\boldsymbol{I})^{p/2}\right) \\ &= \sum_{i=1}^{\min(N,K)}(\sigma_i^2(\boldsymbol{X})+\gamma)^{p/2}, \end{aligned}$$

The $f_p(\boldsymbol{X})$ is differentiable for $p>0$ and convex for $p\geq 1$. With $\gamma=0$, $f_1(\boldsymbol{X})=\|\boldsymbol{X}\|_*$, which is also known as the Schatten-1 norm [Mohan and Fazel, 2012].

Using the Schatten-1 norm as a smooth approximation to the nuclear norm, our objective function becomes

$$\min_{\boldsymbol{X}\geq 0}\|\boldsymbol{L}-\boldsymbol{X}\|_F^2+\lambda\text{tr}\left((\boldsymbol{X}^T\boldsymbol{X}+\gamma\boldsymbol{I})^{1/2}\right)+\beta\|\boldsymbol{X}\|_{2,1} \tag{2}$$

where $\boldsymbol{I}\in\mathbb{R}^{K\times K}$ denotes the identity matrix. Eq. (2) is a convex optimization problem with respect to $\boldsymbol{X}$.

After obtaining the assignment matrix $\boldsymbol{X}$, we apply the normalize cuts method (Ncut) [Shi and Malik, 2000] on the linear kernel of $\boldsymbol{X}$ to obtain the final clustering results.

4.2 Optimization

We propose an iterative algorithm to solve Eq. (2). We first rewrite Eq. (2) as

$$\begin{aligned} J(\boldsymbol{X}) &= \text{tr}(\boldsymbol{X}^T\boldsymbol{X}-2\boldsymbol{L}^T\boldsymbol{X})+\lambda\text{tr}\left((\boldsymbol{X}^T\boldsymbol{X}+\gamma\boldsymbol{I})^{1/2}\right) \\ &\quad +\beta\sum_{j=1}^{K}\|\boldsymbol{X}_{\cdot j}\|_2-\text{tr}(\boldsymbol{\Delta}\boldsymbol{X}^T), \end{aligned} \tag{3}$$

where Lagrangian multipliers $\boldsymbol{\Delta}\in\mathbb{R}^{N\times K}$ enforce the non-negativity constraints $\boldsymbol{X}\geq 0$.

The derivative of Eq. (3) with respect to $\boldsymbol{X}$ is given as

$$\begin{aligned} \frac{\partial \boldsymbol{J}(\boldsymbol{X})}{\partial \boldsymbol{X}} &= 2\boldsymbol{X}-2\boldsymbol{L}+\lambda\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X}+\gamma\boldsymbol{I})^{-1/2}+\beta\boldsymbol{X}\boldsymbol{D}-\boldsymbol{\Delta} \\ &= 2\boldsymbol{X}+\beta\boldsymbol{X}\boldsymbol{D}+\lambda\boldsymbol{X}\boldsymbol{H}^+-\lambda\boldsymbol{X}\boldsymbol{H}^--2\boldsymbol{L}-\boldsymbol{\Delta}, \end{aligned}$$

where $\boldsymbol{D}_{ii}=\frac{1}{\|\boldsymbol{X}_{\cdot i}\|_2}$[1], $\boldsymbol{H}^+=(|\boldsymbol{H}|+\boldsymbol{H})/2$, $\boldsymbol{H}^-=(|\boldsymbol{H}|-\boldsymbol{H})/2$, and $\boldsymbol{H}=(\boldsymbol{X}^T\boldsymbol{X}+\gamma\boldsymbol{I})^{-1/2}$. From the Karush–Kuhn–Tucker (KKT) conditions, $[\frac{\partial \boldsymbol{J}(\boldsymbol{X})}{\partial \boldsymbol{X}}]_{ij}\boldsymbol{X}_{ij}=0$ and $\boldsymbol{\Delta}_{ij}\boldsymbol{X}_{ij}=0,\forall i,j$, we have

$$[2\boldsymbol{X}+\beta\boldsymbol{X}\boldsymbol{D}+\lambda\boldsymbol{X}\boldsymbol{H}^+-\lambda\boldsymbol{X}\boldsymbol{H}^--2\boldsymbol{L}]_{ij}\boldsymbol{X}_{ij}=0. \tag{4}$$

This is a fixed point equation that the solution must satisfy when converged.

Next, we show that the following updating rule satisfies the KKT condition of Eq. (4):

$$\boldsymbol{X}_{ij}^{(t+1)}\leftarrow\boldsymbol{X}_{ij}^{(t)}\sqrt{\frac{[2\boldsymbol{L}+\lambda\boldsymbol{X}^{(t)}\boldsymbol{H}^{-(t)}]_{ij}}{[2\boldsymbol{X}^{(t)}+\beta\boldsymbol{X}^{(t)}\boldsymbol{D}^{(t)}+\lambda\boldsymbol{X}^{(t)}\boldsymbol{H}^{+(t)}]_{ij}}}. \tag{5}$$

The limiting solution of updating rule of Eq. (5) makes the rule to satisfy the fixed point equation: when converged, $\boldsymbol{X}^{(\infty)}=\boldsymbol{X}^{(t+1)}=\boldsymbol{X}^{(t)}=\boldsymbol{X}$ where $t\to\infty$. Then, the updating rule of Eq. (5) reduces

$$\boldsymbol{X}_{ij}\leftarrow\boldsymbol{X}_{ij}\sqrt{\frac{[2\boldsymbol{L}+\lambda\boldsymbol{X}\boldsymbol{H}^-]_{ij}}{[2\boldsymbol{X}+\beta\boldsymbol{X}\boldsymbol{D}+\lambda\boldsymbol{X}\boldsymbol{H}^+]_{ij}}}. \tag{6}$$

From Eq. (6), we get the following equations:

$$[2\boldsymbol{X}+\beta\boldsymbol{X}\boldsymbol{D}+\lambda\boldsymbol{X}\boldsymbol{H}^+-\lambda\boldsymbol{X}\boldsymbol{H}^--2\boldsymbol{L}]_{ij}\boldsymbol{X}_{ij}^2=0. \tag{7}$$

Eq. (7) is identical to Eq. (4). Hence if Eq. (7) holds, Eq. (4) also holds and vice versa. So we have proved that the limiting solution of the updating rule of Eq. (5) satisfies the KKT condition.

The pseudo-code of our optimization process is presented in Algorithm 1.

Complexity Analysis: The complexity of computing $\boldsymbol{H}$ is $\mathcal{O}(K^3)$, because of matrix inversion. Thus, the entire complexity for updating $\boldsymbol{X}$ is $\mathcal{O}(T(K^3+NK^2))$, while the time complexity of INA is $\mathcal{O}(c^2NKT)$ (c and T are the number of clusters and iterations, respectively.). In reality, since K is the order of hundreds and $K\ll N$, the complexity of the proposed method is practically feasible.

[1]When $\|\boldsymbol{X}_{\cdot i}\|_2=0$, we cannot set $\boldsymbol{D}_{ii}$. Thus we use $(\|\boldsymbol{X}_{\cdot i}\|_2+\varsigma)$ which approximates $\|\boldsymbol{X}_{\cdot i}\|_2$, when $\varsigma\to 0$.

Algorithm 1 The RCEC algorithm

Input: $\boldsymbol{S} \in \mathbb{R}^{d \times N}$, the number of clusters c
Output: $\boldsymbol{X} \in \mathbb{R}^{N \times K}$, $consensus\ \ partition$

1: Generate input partitions $\boldsymbol{P}$ from $\boldsymbol{S}$ using clustering algorithms, and construct the binary assignment matrix $\boldsymbol{L}$.
2: Initialize $\boldsymbol{X}^{(0)}$ randomly.
3: **repeat**
4: Compute $\boldsymbol{D}^{(t)} = \frac{1}{\|\boldsymbol{X}^{(t)}\|_2}$
$\boldsymbol{H}^{(t)} = (\boldsymbol{X}^{T(t)}\boldsymbol{X}^{(t)} + \gamma \boldsymbol{I})^{-1/2}$
5: Optimize $\boldsymbol{X}$ by solving Eq. (5).
6: **until converges**
7: Return the final clustering result as :
$consensus\ \ partition = Ncut(\boldsymbol{X}\boldsymbol{X}^T, c)$

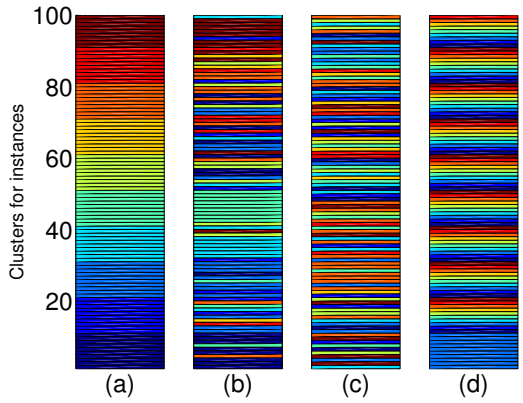


Figure 2: Examples of input partitions (10 clusters). (a) True cluster assignments; (b) Normal partition; (c) Anomalous partition; (d) Extremely anomalous partition. Each color corresponds to a cluster assignment.

5 Experiments

In this section, we first examine the proposed method using synthetic data sets, and then compare the performance of RCEC with five ensemble clustering algorithms HBGF, ECMC, NMFC, WCC, and INA using real-world data sets. For a fair comparison, the true number of clusters is given as a priori parameter to all methods. The performance is evaluated by normalized mutual information (NMI), a widely used information theoretic measure for evaluating clustering methods [Ana and Jain, 2003; Chen and Cai, 2011].

5.1 Synthetic Dataset

In this experiment, we first generate a true partition vector $\boldsymbol{p}^* \in \mathbb{R}^{100}$, which consists of 10 clusters, and each cluster has 10 instances (Figure 2 (a))[2]. Then, we generate 10 normal partitions $\{\boldsymbol{p}_m\}_{m=1}^{10}$ based on the true partition $\boldsymbol{p}^*$ (Figure 2 (b)). A normal partition $\boldsymbol{p}_m$ consists of 5 normal clusters and 5 anomalous clusters. In a normal cluster, only 20% of instances are randomly permuted to other clusters, while in an anomalous cluster, 80% of instances are randomly permuted to other clusters. Moreover, we prepare some anomalous partitions and extremely anomalous partitions. As shown in Figure 2 (c), each anomalous partition $\boldsymbol{p}'_m$ consists of 10 anomalous clusters. We use the pooled normal and anomalous partitions $\{\boldsymbol{p}_m\}_{m=1}^{10} \cup \{\boldsymbol{p}'_m\}_{m=1}^{n'}$ as the input. We vary the number of (extremely) anomalous partitions n' and report the average NMI score. Note that the extremely anomalous partition (Figure 2 (d)) contains one true cluster and nine extremely anomalous clusters, where all instances scatter in different true clusters.

[2]In Figure 2, The y-axis indicates instances, and different colors correspond to different clusters. The normal partition includes more normal clusters than the anomalous partitions which consist of a few normal clusters and many anomalous clusters.

Table 1: Data sets used in the experiments . N is the number of instances, c is the number of clusters, d is the number of feature and N_LC (N_SC) is the number of instances in the largest (smallest) cluster.

Datasets	N	c	d	N_LC	N_SC
Tr11 [Karypis, 2002]	414	9	6429	132	6
K1b [Karypis, 2002]	2340	6	13879	1389	60
ORL [Cai *et al.*, 2006]	400	40	1024	10	10

Figure 3 (a) and (b) show the results of Ave[3], HBGF, ECMC, NMFC, WCC, INA and RCEC when adding anomalous partitions and extremely anomalous partitions, respectively. As shown in these figures, the performance of existing methods significantly decreased by adding anomalous clusters, while RCEC was more robust against anomalous clusters.

Furthermore, we checked whether the RCEC algorithm can safely get rid of anomalous clusters. To this end, we define two indicators (normalized magnitude) $l_j / \max_j(l_j)$ and $x_j / \max_j(x_j)$, where $l_j = \sum_{i=1}^{N} \boldsymbol{L}_{ij}, x_j = \sum_{i=1}^{N} \boldsymbol{X}_{ij}, j = 1, 2, \ldots, K$. Please note when $l_j / \max_j(l_j)$ is large and $x_j / \max_j(x_j)$ is small (possibly zero), the cluster would be likely to be an anomalous cluster. Figure 4 (a) and (b) show the indicators of $\boldsymbol{L}$ and $\boldsymbol{X}$ when $n' = 5$ and extremely anomalous partitions are added. From the figures, we can see that clusters 101 to 150 correspond to extremely anomalous clusters, and blue and red regions correspond to normal and anomalous clusters, respectively. We can then clearly see that RCEC could retain normal clusters (blue) and filter out anomalous clusters (red).

5.2 Real-world Data sets

We evaluated the RCEC algorithm using two text and one image data sets, which are summarized in Table 1.

For the text data sets, we used the TFIDF [Manning *et al.*, 2008] features. We randomly chose $60\%, 70\%, \ldots, 100\%$ of the entire features for experiments. In this experiment, 30 partitions estimated by k-means were used as the input. We repeated the experiment 10 times by changing the random seed and reported the average NMI values. We used $\lambda = 0.1$, $\gamma = 0.01$, and $\beta = \{0.01, 1, 2, 4, 6, \ldots, 20\}$. For NMF-based

[3]"Ave" is the average NMI values of all input partitions (clustering algorithms).

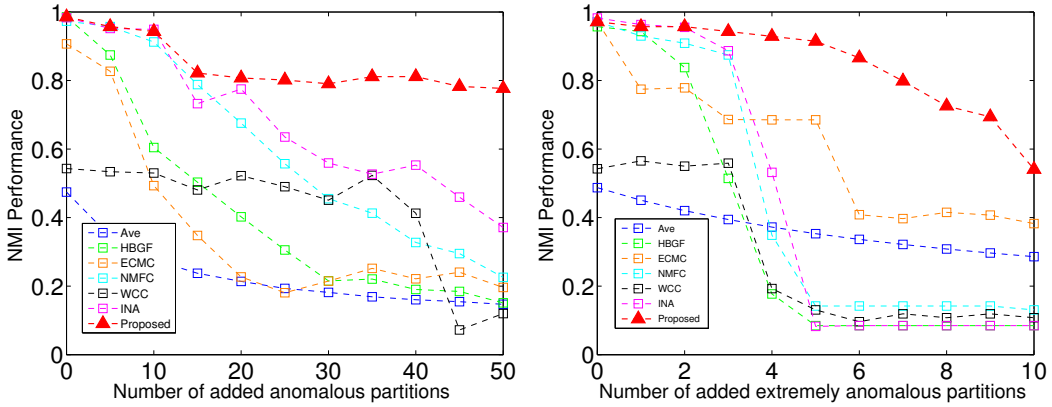


Figure 3: Performance comparison on the synthetic data sets when (a) anomalous partitions and (b) extremely anomalous partitions are added.

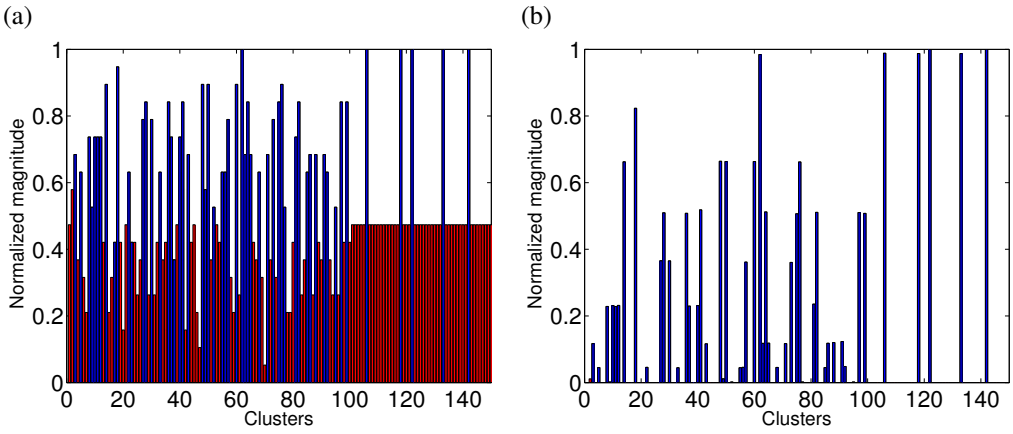


Figure 4: Illustration of the RCEC algorithm. Clusters 1 to 100 correspond to 10 normal input partitions, and clusters 101 to 150 represent 5 extremely anomalous input partitions. Blue regions are normal clusters, and red regions are anomalous clusters. (a) The normalized magnitude of the input matrix $\boldsymbol{L}$; (b) The normalized magnitude of the learned matrix $\boldsymbol{X}$.

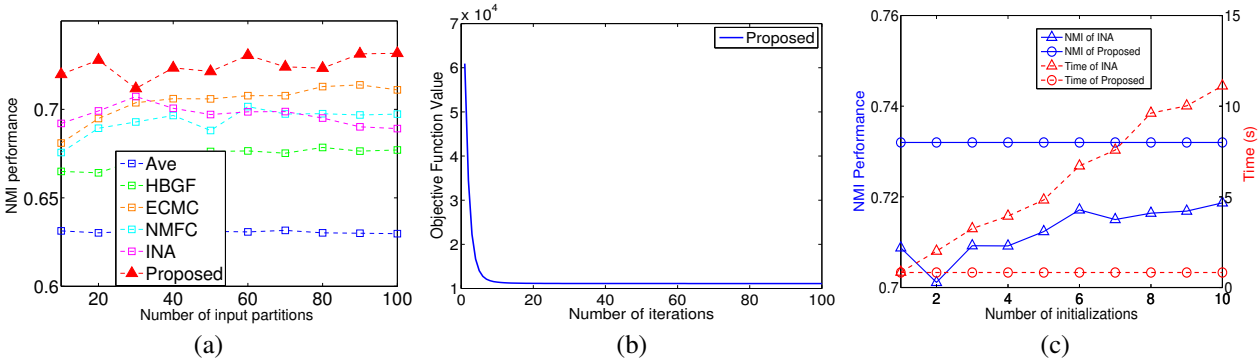
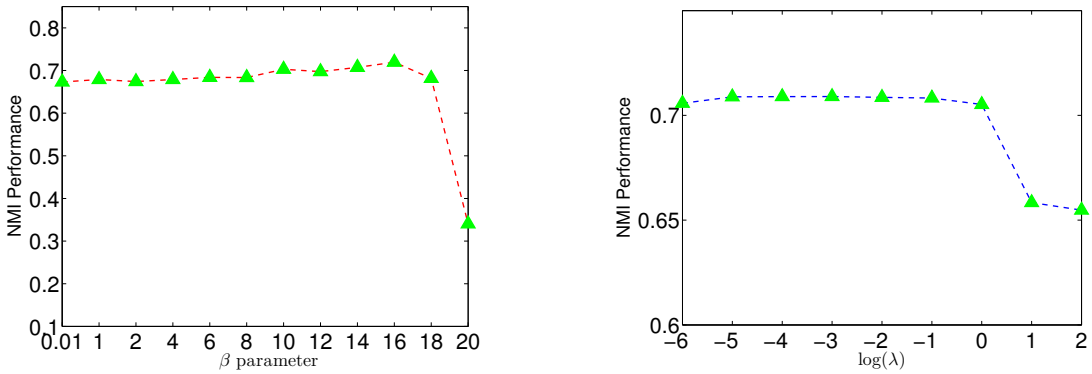


Figure 5: Results on the real-world data sets. (a) Performance with changing the number of input partitions on the Tr11 data set; (b) Objective function value of RCEC with respect to the number of iterations on the Tr11 data set; (c) Performance and computation time of INA and RCEC with respect to the number of initializations.



(a) (b)

Figure 6: (a) Performance of RCEC with changing parameter β; (b) Performance of RCEC with changing parameter λ.

Table 2: NMI values on the Tr11 data set. We compared the best method having the highest mean NMI with all other methods using paired t-test. The best method at the significance level 5% is highlighted in boldface.

Ratio (%)	Ave	HBGF	ECMC	NMFC	INA	Proposed
60	0.6298	0.6651	0.7002	0.6898	0.7007	**0.7154**
70	0.6408	0.6823	0.7076	0.7041	0.7115	**0.7292**
80	0.6523	0.6808	0.7055	0.6988	0.7057	**0.7299**
90	0.6589	0.6790	0.7066	0.7055	0.7117	**0.7225**
100	0.6674	0.7002	0.7147	0.7120	0.7150	**0.7307**

Table 3: NMI values on the K1b data set.

Ratio (%)	Ave	HBGF	ECMC	NMFC	INA	Proposed
60	0.5916	0.6315	0.6478	0.6159	0.6675	**0.6735**
70	0.5994	0.6346	0.6532	0.6129	0.6723	**0.6764**
80	0.5948	0.6347	0.6511	0.6083	0.6617	**0.6686**
90	0.6016	0.6402	0.6653	0.6003	0.6620	**0.6714**
100	0.6061	0.6560	0.6687	0.6118	0.6697	**0.6813**

Table 4: NMI values on the ORL data set.

Ratio (%)	Ave	HBGF	ECMC	NMFC	INA	Proposed
60	0.7560	0.7743	0.6825	0.7776	0.7588	**0.7812**
70	0.7538	0.7733	0.6815	0.7790	0.7612	**0.7830**
80	0.7577	0.7747	0.6835	0.7806	0.7685	**0.7888**
90	0.7565	0.7737	0.6960	0.7818	0.7717	**0.7853**
100	0.7600	0.7786	0.6980	0.7893	0.7695	0.7881

ensemble methods (including NMFC, WCC, and INA), we ran those algorithms five times with different initial parameter values and averaged over the clustering results. This is a common heuristic to stabilize the ensemble clustering algorithms.

Tables 2, 3 and 4 show the performance of different algorithms in terms of NMI. We could not include WCC, since WCC is computationally too expensive. The three tables clearly indicate that RCEC outperformed existing methods, being statistically significant.

Number of Partitions and Parameters: We evaluated the clustering performance as a function of the number of input partitions. The process of generating input partitions was repeated 10 times. In the experiments, we used 60% of input features, $\gamma = 0.01$, and $\lambda = 0.1$. Figure 5 (a) shows the results with changing the number of input partitions on the Tr11 data set. RCEC outperformed other methods through all the numbers of input partitions.

RCEC has two essential parameters β and λ. Figure 6 shows how the performance of RCEC varied with the parameter values of β and λ. From the results, the performance of RCEC could be kept consistently when β is within $[10, 16]$ and λ in $[10^{-6}, 10^{0}]$.

Convergence and Computation: Figure 5 (b) shows the speed of convergence of the proposed algorithm on the Tr11 data set. We can see that the multiplicative updating rule converged very quickly, usually within 20 iterations. We also compared the computation time of the proposed method with INA (Figure 5 (c)). In this experiment, since INA is a non-convex method, we ran INA several times with different initializations and averaged over those clustering results to obtain a stable result. From the results, we can see that the clustering performance of INA were better by increasing a larger number of initializations, which however results in a larger amount of computation time. On the other hand, RCEC can find the global minimum, by which we just need to run RCEC only once. Hence RCEC runs faster than INA in practice, which makes the performance of RCEC higher.

6 Conclusion

We have proposed a new method for ensemble clustering which we call Robust Convex Ensemble Clustering (RCEC). The key idea of RCEC is to use the nuclear norm to handle the low-rank structure of the cluster assignment matrix and the group norm to detect anomalous clusters. The formulation has the convex property, which allows to obtain the global minimum, for which we have presented a simple yet effective multiplicative updating rule. Extensive experiments on synthetic and real data sets showed the effectiveness and efficiency of our approach comparing to the state-of-the-art ensemble clustering methods.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61572139) and ICR-KU International Short term Exchange Program for Young Researchers. Funding acknowledgement: M.Y.: MEXT KAKENHI #16K16114, S.K.: Academy of Finland, COIN and grants 294238 and 292334, H.M.: MEXT KAKENHI #16H02868, M.Y. and H.M.: JST, ACCEL, "Reinforcement of Resiliency of Concentrated Polymer Brushes and Its Tribological Applications", H.M. and S.K.: Tekes, FiDiPro.

References

[Ana and Jain, 2003] LNF Ana and Anil K Jain. Robust data clustering. In *CVPR*, 2003.

[Cai *et al.*, 2006] Deng Cai, Xiaofei He, Jiawei Han, and Hong-Jiang Zhang. Orthogonal laplacianfaces for face recognition. *IEEE Trans. on Image Processing*, 15(11):3608–3614, 2006.

[Chen and Cai, 2011] Xinlei Chen and Deng Cai. Large scale spectral clustering with landmark-based representation. In *AAAI*, 2011.

[Ding *et al.*, 2006] Chris Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *KDD*, 2006.

[Fern and Brodley, 2004] Xiaoli Zhang Fern and Carla E Brodley. Solving cluster ensemble problems by bipartite graph partitioning. In *ICML*, 2004.

[Hartigan and Wong, 1979] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Applied statistics*, pages 100–108, 1979.

[He *et al.*, 2012] Ran He, Tieniu Tan, Liang Wang, and Wei-Shi Zheng. $\ell_{2,1}$ regularized correntropy for robust feature selection. In *CVPR*, 2012.

[Huang *et al.*, 2011] Xiaodi Huang, Xiaodong Zheng, Wei Yuan, Fei Wang, and Shanfeng Zhu. Enhanced clustering of biomedical documents using ensemble non-negative matrix factorization. *Information Sciences*, 181(11):2293–2302, 2011.

[Jain, 2010] Anil K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recogn. Lett.*, 31(8):651–666, June 2010.

[Karypis, 2002] George Karypis. Cluto-a clustering toolkit. Technical report, DTIC Document, 2002.

[Li and Ding, 2008] Tao Li and Chris Ding. Weighted consensus clustering. *SDM*, 2008.

[Li *et al.*, 2007] Tao Li, Chris Ding, Michael Jordan, et al. Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In *ICDM*, 2007.

[Manning *et al.*, 2008] Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[Mohan and Fazel, 2012] Karthik Mohan and Maryam Fazel. Iterative reweighted algorithms for matrix rank minimization. *Journal of Machine Learning Research*, 13(1):3441–3473, 2012.

[Ng *et al.*, 2002] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. *NIPS*, 2002.

[Shi and Malik, 2000] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 22(8):888–905, 2000.

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

[Topchy *et al.*, 2005] Alexander P. Topchy, Anil K. Jain, and William F. Punch. Clustering ensembles: Models of consensus and weak partitions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(12):1866–1881, 2005.

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

[Yang *et al.*, 2011] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou. l2,1-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI*, 2011.

[Yang *et al.*, 2013] Shizhun Yang, Chenping Hou, Changshui Zhang, and Yi Wu. Robust non-negative matrix factorization via joint sparse and graph regularization for transfer learning. *Neural Computing and Applications*, 23(2):541–559, 2013.

[Yi *et al.*, 2012] Jinfeng Yi, Tianbao Yang, Rong Jin, Anubhav K Jain, and Mehdi Mahdavi. Robust ensemble clustering by matrix completion. In *ICDM*, 2012.

[Yuan and Lin, 2006] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[Zheng *et al.*, 2015] Xiaodong Zheng, Shanfeng Zhu, Junning Gao, and Hiroshi Mamitsuka. Instance-wise weighted nonnegative matrix factorization for aggregating partitions with locally reliable clusters. In *IJCAI*, 2015.

[Zhou *et al.*, 2015] Peng Zhou, Liang Du, Hanmo Wang, Lei Shi, and Yi-Dong Shen. Learning a robust consensus matrix for clustering ensemble via kullback-leibler divergence minimization. In *IJCAI*, 2015.

[Zhou, 2012] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC Press, 2012.