Ensemble Clustering using Semidefinite Programming

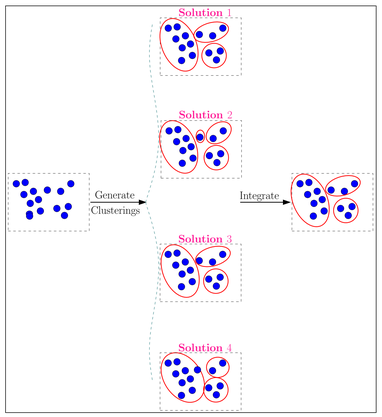
Vikas Singh[1], Lopamudra Mukherjee[2], Jiming Peng[3], Jinhui Xu[2]

[1]Biostatistics and Medical Informatics
University of Wisconsin – Madison
vsingh@biostat.wisc.edu

[2]Computer Science & Engineering
State University of New York at Buffalo
{lm37,jinhui}@cse.buffalo.edu

[3]Industrial and Enterprise Systems Eng.
University of Illinois Urbana Champaign
pengj@uiuc.edu

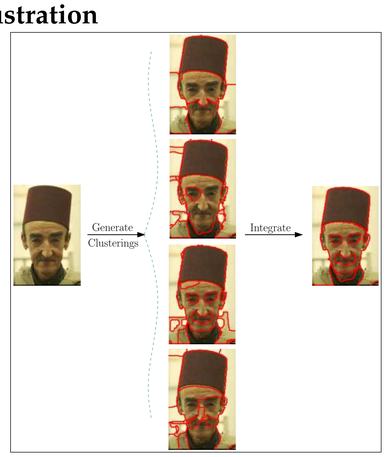
The Problem:
How to best combine the **ensemble of multiple clustering solutions** in a maximum consent sense?



Motivation

1. No single clustering algorithm is perfect.
2. No ground truth is known in many applications.
3. Ensemble likely to be superior to individual solutions.

An Illustration



Problem Statement

Given: $P_1, P_2, \ldots, P_m$ be m partitions of the data, each P_i is produced by a different algorithm, C_i
Determine: A partition P^* for the ensemble that maximizes 'agreement' between C_i's.

Previous Works

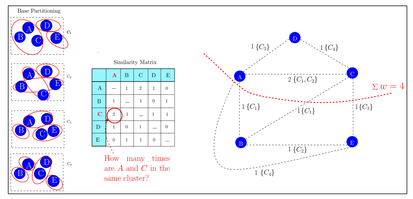
- Idea first proposed by Strehl and Ghosh.
- Existing techniques are **primarily Graph based Formulation**, details below.
- Relevant papers include Strehl and Ghosh (AAAI, JMLR 2002), Fern and Brodley (ICML 2003).

Based on . . .

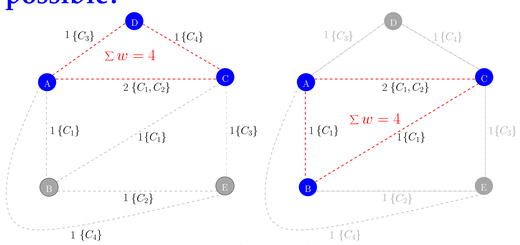
1. Count how many times algorithms put a pair of items together.
2. Construct graph with clusters or items as nodes.
3. Assign weights (i.e., togetherness frequency) to edges between node-pairs.
4. Do graph partitioning.

In other variants, each cluster is a node and edge weights are evaluated by evaluating pairs of clusters.

An illustration of IBGF

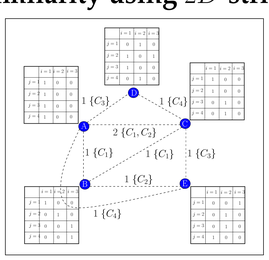


Are pairwise voting strategies the best possible?



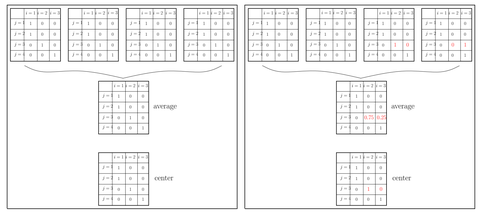
Which algorithms put (A, C, D) together? None.
Which algorithms put (A, B, C) together? At least C_1.

Modeling similarity using $2D$-strings



Averaging strings to calculate similarity

If solutions were known, we could . . .



Optimization Problem

Given $D = (d_1, \ldots, d_n)$ and ensemble, $C = (C_1, \ldots, C_m)$, partition the items, i.e., A-strings into k clusters, s.t. *the absolute difference of the average of clusters to their centers is minimized*

$$\min \sum_{i'=1}^{k}\sum_{i=1}^{k}\sum_{j=1}^{m} \left\| \underbrace{s_{i'ij}}_{center} - \frac{\sum_{l=1}^{n} A_{lij} X_{li'}}{\underbrace{\sum_{l=1}^{n} X_{li'}}_{average}} \right\|$$

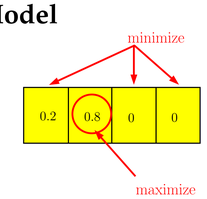
$$s.t \sum_{\overline{i'=1}}^{k} X_{li'} = 1;\ \forall l \in [1, n]$$

$$\sum_{l=1}^{k} X_{li'} \geq 1;\ \forall i' \in [1, k]$$

$$\sum_{i=1} s_{i'ij} = 1;\ \forall j \in [1, m], \forall i' \in [1, k].$$

Note: Model can be converted to a strict 0-1 SDP, but all variables (especially s) cannot be relaxed.

Alternative Model



Replace $1 - \max(a_1, a_2, \ldots, a_k)$ with $1 - ({a_1}^2 + \ldots + {a_k}^2)$?

$$\min \sum_{i'=1}^{k}\sum_{j=1}^{m}\Big(\sum_{l=1}^{n} X_{li'}\Big)\left(1 - \sum_{i=1}^{k}\left(\frac{\sum_{l=1}^{n} A_{lij} X_{li'}}{\sum_{l=1}^{n} X_{li'}}\right)^2\right)$$

$$s.t \sum_{\overline{i'=1}}^{k} X_{li'} = 1 \quad \forall l \in [1, n]$$

$$\sum_{l=1} X_{li'} \geq 1 \quad \forall i' \in [1, k].$$

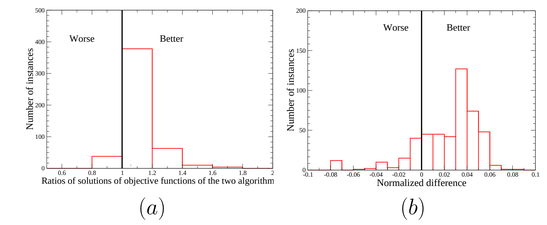
Equivalent SDP

$$\min \ (nm - \mathrm{tr}(BZ))$$
$$s.t. \ Z_{i'}\mathbf{1} \leq \mathbf{1} \quad \forall i' \in [1, k]; \mathrm{tr}(Z) = k,$$
$$Z \geq 0; Z^2 = Z; Z = Z^T.$$

where $B = \sum_{j=1}^{m}\sum_{i=1}^{k} A_{ij}^T A_{ij}$ and $Z = X(X^TX)^{-1}X^T$.

Experimental Results

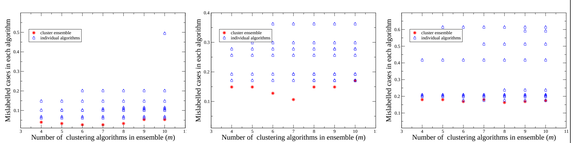
Comparison results with Strehl and Ghosh



(a) (b)

Explanation. A comparison of Strehl & Ghosh with SDP Model 2 in (a). In (b), comparisons (difference in normalized values) between our solution and the best among two algorithms with the Normalized Mutual Information (NMI) objective function used in Strehl & Ghosh .

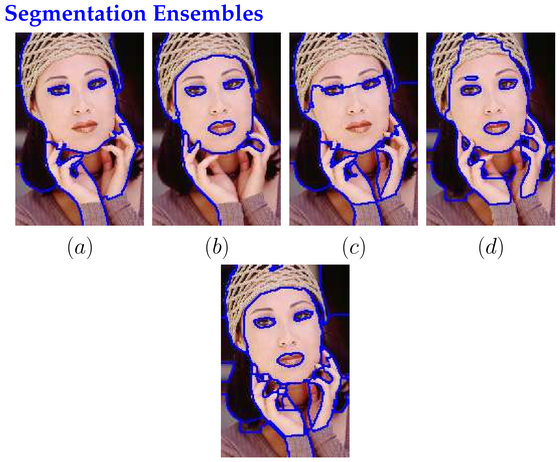
UCI datasets



(a) (b) (c)

Explanation. The fraction of mislabeled cases in a consensus solution ($*$) is compared to the number of mislabelled cases(δ) in individual clustering algorithms for the Iris dataset in (a), the Soybean dataset in (b), and the Wine dataset in (c).

Segmentation Ensembles



(a) (b) (c) (d) (e)

Explanation. (a)–(d) show the individual segmentations overlaid on the input image, (e) shows the segmentation generated from ensemble clustering.

Advances in Neural Information Processing Systems, 2007.