Exploring the Performance Limit of Cluster Ensemble Techniques

Xiaoyi Jiang and Daniel Abdala
Department of Mathematics and Computer Science
University of Münster
Germany

Cluster ensemble as optimization problem

- Data set X = {x_1, x_2, …, x_n} of n patterns
- A cluster ensemble is a set P = {P_1, P_2, …, P_N}, where P_i is a clustering of X
- Denote the set of all possible clusterings of X by P_X
- Cluster ensemble as median partition problem:

$$P^* = \arg \min_{P \in \mathcal{P}_X} \sum_{i=1}^{N} d(P, P_i)$$

d(): distance (dissimilarity) function between two clusterings

Cluster ensemble as optimization problem

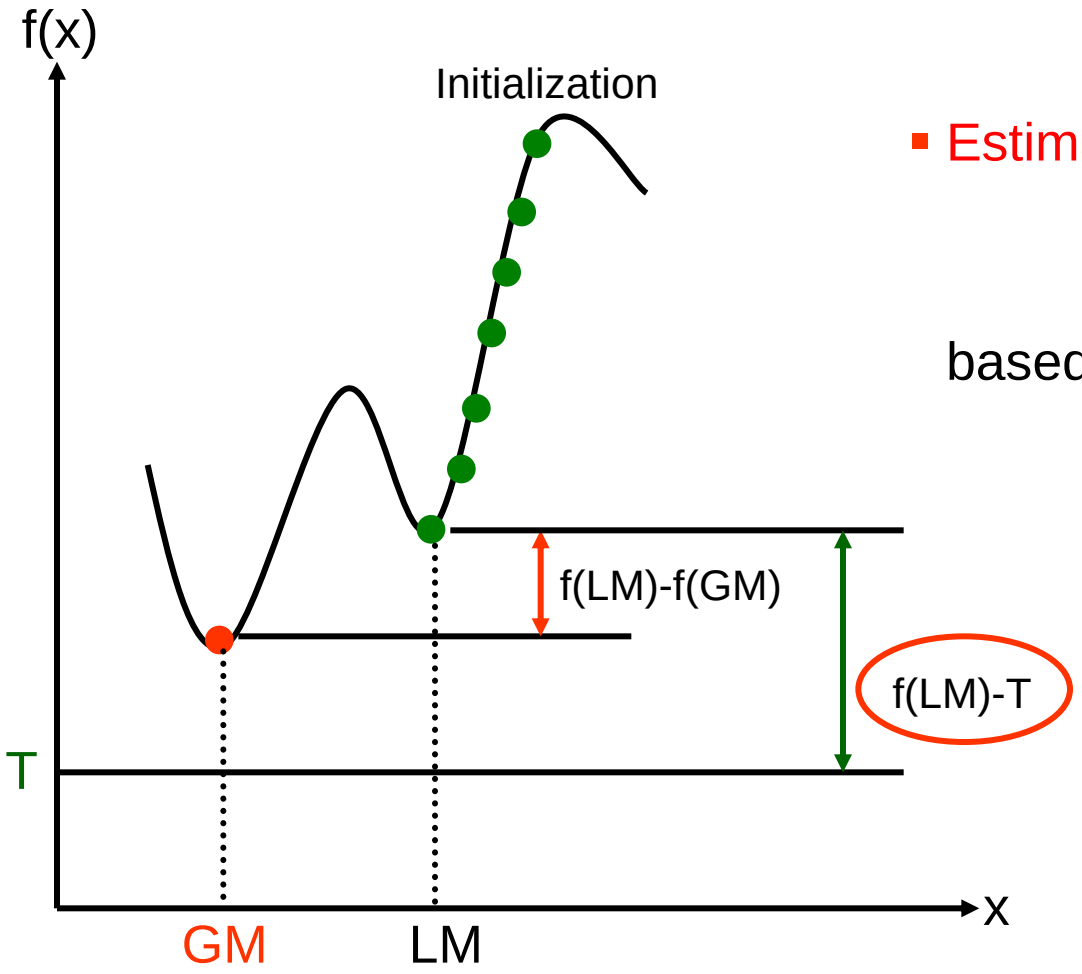
- Cluster ensemble as median partition problem:

$$P^* = \arg \min_{P \in \mathcal{P}_X} \sum_{i=1}^{N} d(P, P_i)$$

- This is a difficult optimization problem
  - ✓ Simple solution only for trivial distance functions
  - ✓ For several distance functions it turns out to be NP-complete (seems to be proved that P != NP)
  - ✓ Also for other distance functions no efficient solutions are known

→ Approximate methods → How good is the suboptimal solution?

Illustratation: How good is a local minimum?



- Goodness of local minimum LM

  f(LM) – f(GM)

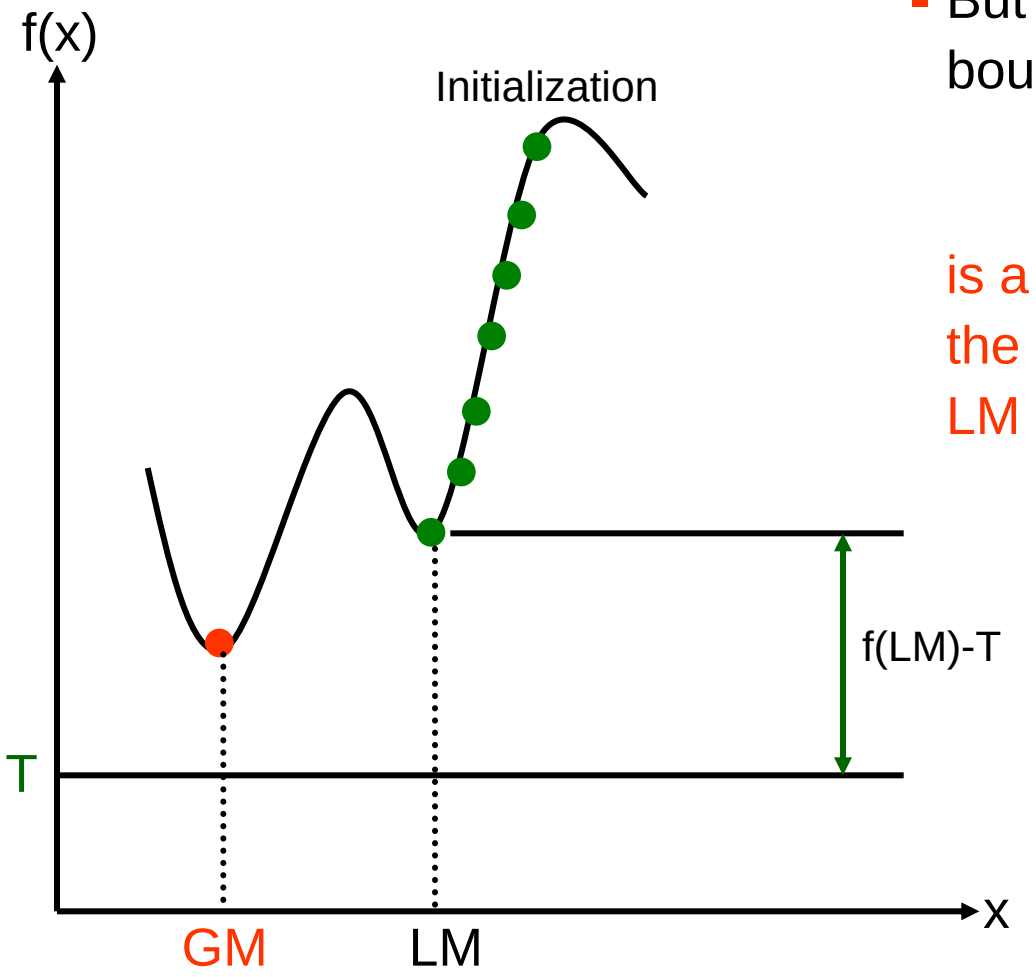
- Estimation of goodness

  f(LM) – T

  based on lower bound T with

  f(x) >= T

Illustratation: How good is a local minimum?



- Lower bound T = 0 not useful
- But given some tight lower bound T

$$f(LM) - T$$

is a good option for measuring the goodness of local minimum LM

Lower bound für median partition problem

- Median partition problem is an instance of generalized median problem (appplicable to any object domains)
- Lower bound T for generalized median problem in metric space based on linear programming (Jiang & Bunke, SSPR 2002):

$$\text{minimize } x_1 + x_2 + \cdots + x_N \text{ subject to}$$

$$\forall i, j \in \{1, 2, \ldots, N\},\ i \neq j,\ \begin{cases} x_i + x_j \geq d(P_i, P_j) \\ x_i + d(P_i, P_j) \geq x_j \\ x_j + d(P_i, P_j) \geq x_i \end{cases}$$

$$\forall i \in \{1, 2, \ldots, N\},\ x_i \geq 0$$

- It turns out to be tight in several other contexts

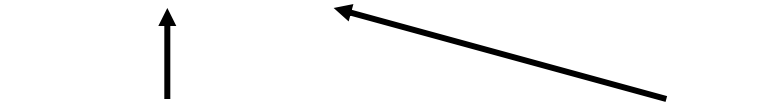
Experimental results

- Metric distance functions
  - ✓ Variance of information
  - ✓ van Dongen metric
  - ✓ Mirkin metric
- Cluster ensemble methods
  - ✓ Evidence accumulation method (Fred & Jain, PAMI 2005)
  - ✓ Random walker based method (Abdala et al. ICPR 2010)
- Data sets
  - ✓ Nine UCI data sets
  - ✓ Two articial data sets

Experimental results

Random walker based method `RW`

	d_{vi}			d_{vd}			d_m		
dataset	SOD($\tilde{P}$)	Γ	$\Delta'(\%)$	SOD($\tilde{P}$)	Γ	$\Delta'(\%)$	SOD($\tilde{P}$)	Γ	$\Delta'(\%)$
Iris	8.40	7.24	13.8	2.28	2.16	5.2	28067	25113	10.5
Wine	2.09	1.86	10.0	0.35	0.33	4.5	7242	6777	5.8
Breast	1.49	1.08	27.7	0.20	0.15	23.9	90032	68392	24.0
Optic	11.38	6.37	44.0	3.90	1.85	50.9	749459	315016	57.7
Soy	6.19	3.79	36.9	4.08	1.62	52.0	3433	1591	49.3
Glass	7.96	4.66	41.1	2.53	1.24	45.9	69186	33940	49.3
Haberman	7.70	7.58	1.5	2.86	2.84	0.7	234484	232995	0.6
Mammo	1.77	1.77	0.0	0.38	0.38	0.0	248650	248650	0.0
Yeast	18.60	11.40	38.2	10.51	3.34	67.5	6606869	3010185	53.4
2D2K	4.69	4.69	0.0	1.15	1.15	0.0	978050	978050	0.0
8D5K	5.24	4.91	5.9	2.43	1.66	15.0	721412	579262	11.3



Experimental results

- For three data sets (Haberman, Mammo, and 2D2K) the lower bound T is almost reached for all three distance functions
  → practically no room for improvement

- If the deviation is large, we must be careful in making any claims
  - ✓ The lower bound is not tight enough in that particular case
  - ✓ The computed solution is still far away from the (unknown) optimal solution

Larger deviation may indicate some, although uncertain, potential of improvement and thus serves as a hint for continuing optimization.

Additional contents of the paper

- Merkin distance: General lower bound T is almost as good as with domain-specific lower bound

- Extension to weighted cluster ensemble techniques

$$P^* = \arg \min_{P \in \mathcal{P}_X} \sum_{i=1}^{N} w_i \cdot d(P, P_i)$$

- Extension to quasi-metric distance functions

$$d(P, R) + d(R, Q) \geq \frac{d(P, Q)}{1 + \varepsilon}$$

Conclusion: The lower bound T may be considered as a means of exploring the performance limit of cluster ensemble techniques

Experimental results

- What is a good consensus function for ensemble clustering?
- What is a good weighting scheme for a particular consensus function?
- How to assess to which extent a suboptimal algorithm has found a good consensus solution?
- A tighter lower bound for the ensemble clustering problem?
- Lower bound for non-metric distance function?
- Other options for evaluating the quality of a consensus partition