

Objectives Multi-Run Clustering

1. Provide a system that automatically conducts experiments: different clustering algorithm and fitness functions parameters are selected using reinforcement learning, experiments will be run, the promising results will be stored, more experiments will be run, and finally the results are summarized and the best results will be presented to the user.
2. Improve clustering results by using clusters obtained in different runs of a clustering algorithms; the final clustering result will be constructed by decomposing a clustering taking clusters that have been obtained in multiple runs.
3. To support finding clusters that are good with respect to multiple objective (fitness) functions
4. To overcome initialization problems that most clustering algorithms face

Part 3a MRC Scenario

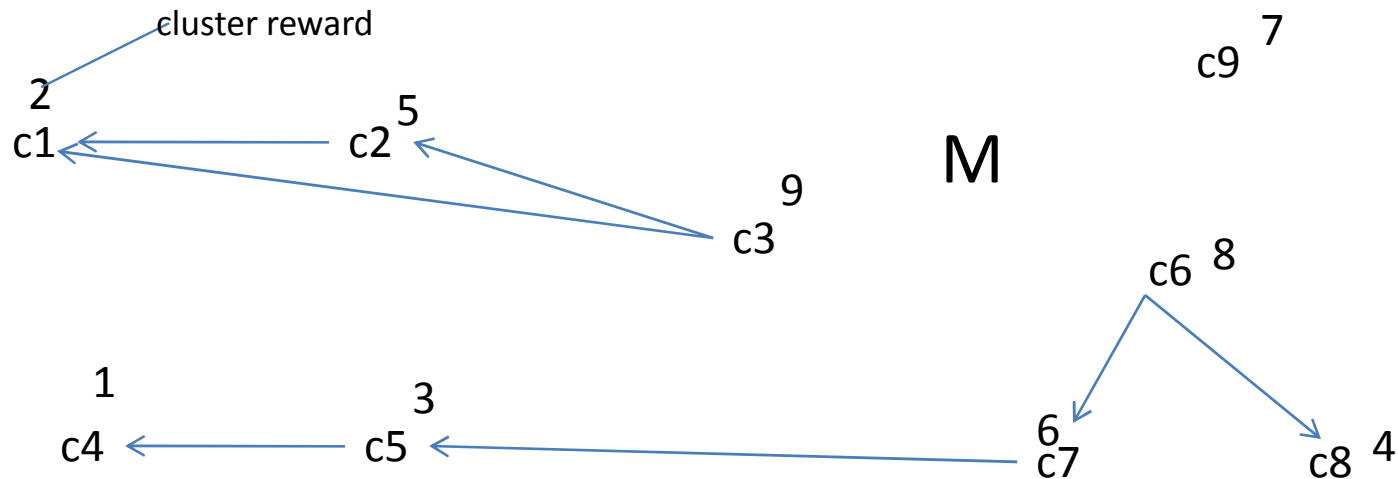
Let c, c' be two clusters

$$\text{sim}(c, c') = \frac{|c \cap c'|}{|c \cup c'|}$$

$$d(c, c') = 1 - \text{sim}(c, c')$$

Challenge: low quality, and high overlap clusters should be removed.

Example: Let us assume we will remove all clusters whose similarity is above 0.3.
 $c \rightarrow c' := \text{reward}(c) > \text{reward}(c')$ and $\text{sim}(c, c') > 0.3$



Ensemble Clustering Problem : what $M' \subset M$ should we return?

Other Thoughts

- Conduct experiments that compare the single-run with multi-run clustering
 - E.g. run CLEVER 20 times with slightly different parameter settings for the Earthquake dataset and compare the multi-run result with the best single run result
- Possible evaluation measures:
 - Cluster reward (likely the most popular choice)
 - Cluster reward per object
- Your system could also be an interactive system

Problem what $M' \subset M$ should we return?

Meta Clustering

- Idea: Cluster the clusters obtained from multiple runs and then present the end user with a summary:
 - e.g user picks the best cluster in each meta cluster
 - System picks best cluster automatically based on cluster reward or cluster reward per object or interestingness
- Distance Measure: $1 - \text{sim}(c, c') = \frac{|c \cap c'|}{|c \cup c'|}$
- Fitness functions
 - MSE (traditional clustering)
 - For variance problem: compare mean-values and variance of cluster
 - For co-location problem: Compare average product of z-scores
 - Use a combination of MSE and some something else
- Again conduct some experiment that demonstrates the benefits of the system you developed.