Comparison of different conflict resolution strategies in collaborative clustering

Germain Forestier

University of Strasbourg
(France)

Image Sciences, Computer Sciences and Remote Sensing Laboratory
–
Groupe de Travail EGC sur la Fouille de Données Complexes (GT EGC-FDC)

28 Juin 2010

Collaborative clustering

- **Aim of this work** : comparison of different conflict resolution strategies in collaborative clustering
- Presentation of a new collaborative clustering approach using a **genetic algorithm**
- Comparison of the methods on several data sets (artificial and UCI)

Collaborative clustering

- There exists **a lot of** different clustering methods (and new ones are published everyday)
- Instead of creating a new one, we tried to make **collaborate** the existing ones
- We proposed a method of **collaborative clustering** which takes advantage of different clustering results
- The method consists in an automatic and mutual **refinement** of different clustering results
- The idea is to produce several **different clustering** results which collaborate to **find an agreement** about the clustering of a dataset

Data

	a_1	a_2
x_1	0.366	2.683
x_2	0.258	2.743
x_3	0.361	2.818
x_4	0.746	2.720
x_5	0.337	2.896
x_6	0.036	2.410
x_7	0.478	2.301
...	...	...

Tab.: The data.

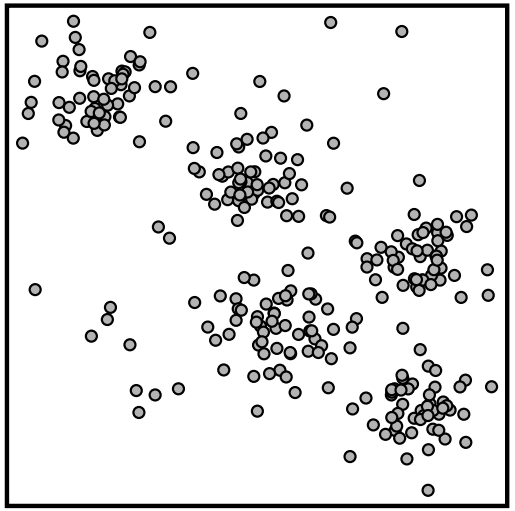


Fig.: The data in a 2dim space.

- Let $X = \{x_1, \ldots, x_n\}$ be the data to process.

Clustering

	a_1	a_2	cluster
x_1	0.366	2.683	c_1
x_2	0.258	2.743	c_2
x_3	0.361	2.818	c_2
x_4	0.746	2.720	c_1
x_5	0.337	2.896	c_3
x_6	0.036	2.410	c_1
x_7	0.478	2.301	c_2
...	...	...	...

Tab.: The data.

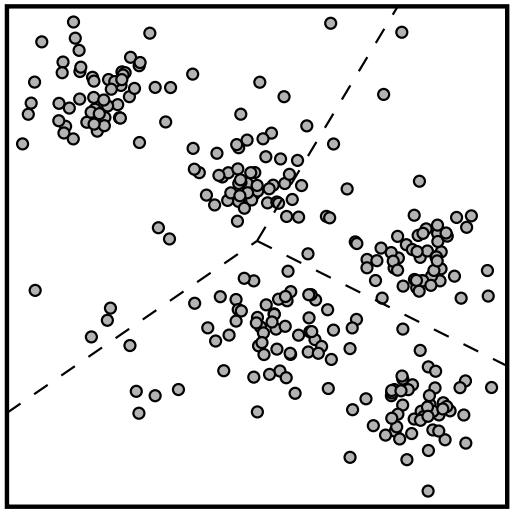
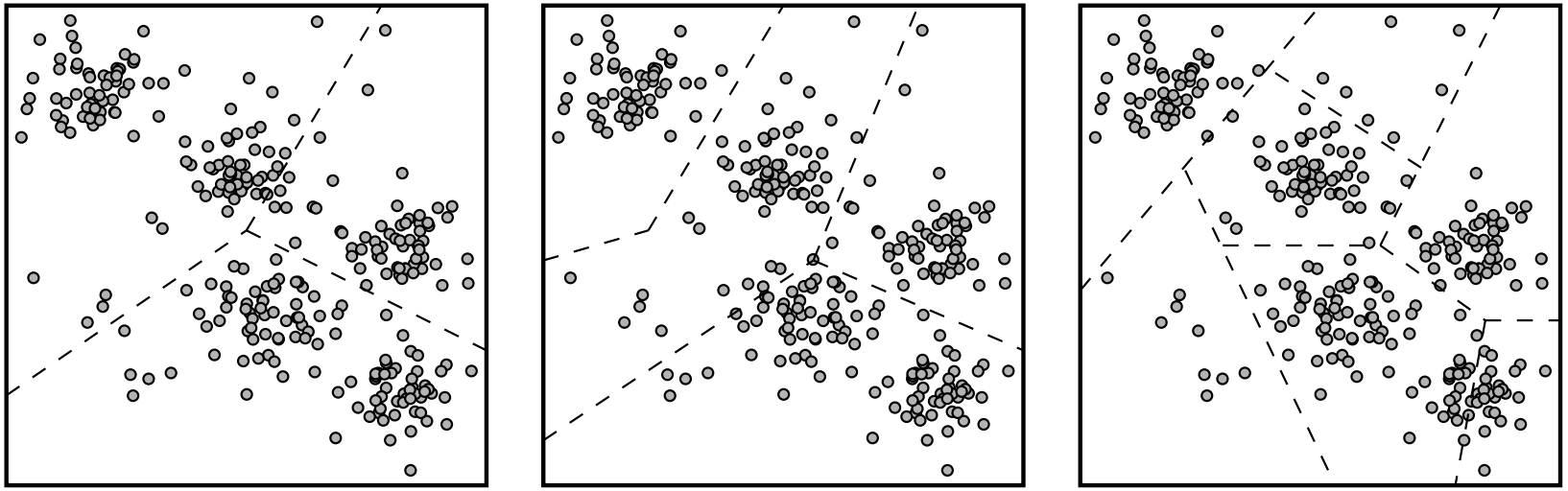


Fig.: Example of one clustering of the data.

- Let $\mathcal{C} = \{C_1, \ldots, C_K\}$ be one clustering result.

Example of three different clustering results



(a) 3 clusters (b) 4 clusters (c) 7 clusters

Fig.: Example of three different clustering results.

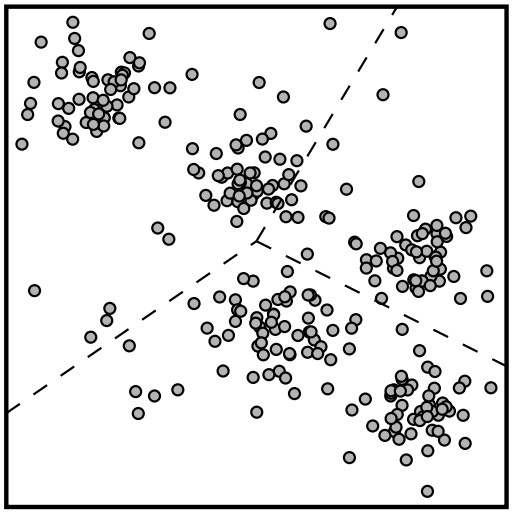
- Let $\mathbb{C} = \{\mathcal{C}^{(1)}, \ldots, \mathcal{C}^{(N)}\}$ be the ensemble of clustering results.
- Let $\mathcal{C}^{(i)} = \{C_1^{(i)}, \ldots, C_{K^{(i)}}^{(i)}\}$ be the ith clustering of the ensemble.

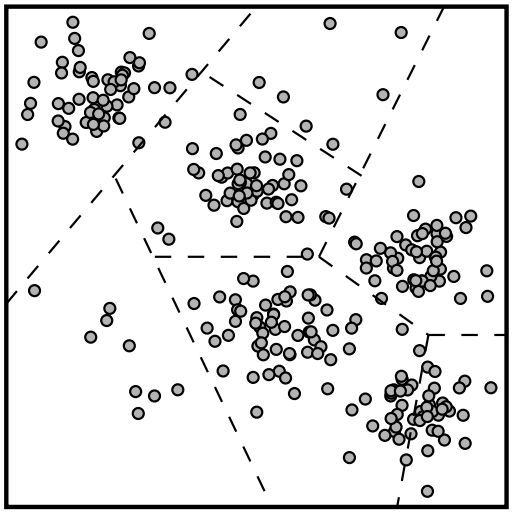
The confusion matrix

- To compute the similarity between the different results, the intersections between each couple of clusters $\left(C_k^{(i)}, C_l^{(j)}\right)$, from two results $\mathcal{C}^{(i)}$ and $\mathcal{C}^{(j)}$, are computed in the **confusion matrix** $\Omega^{(i,j)}$:

$$\Omega^{(i,j)} = \begin{pmatrix} \alpha_{1,1}^{(i,j)} & \cdots & \alpha_{1,K^{(j)}}^{(i,j)} \\ & \vdots & \\ \alpha_{K^{(i)},1}^{(i,j)} & \cdots & \alpha_{K^{(i)},K^{(j)}}^{(i,j)} \end{pmatrix} \text{ where } \alpha_{k,l}^{(i,j)} = \frac{\left|C_k^{(i)} \bigcap C_l^{(j)}\right|}{\left|C_k^{(i)}\right|} \quad (1)$$

The confusion matrix





	$C_1^{(1)}$	$C_2^{(1)}$	$C_3^{(1)}$
$C_1^{(2)}$		0.11	0.42
$C_2^{(2)}$		0.05	0.03
$C_3^{(2)}$			0.50
$C_4^{(2)}$	0,45		
$C_5^{(2)}$	0,05		0,05
$C_6^{(2)}$	0,50	0,02	
$C_7^{(2)}$		0,82	

Tab.: $\Omega^{(1,2)}$

	$C_1^{(2)}$	$C_2^{(2)}$	$C_3^{(2)}$	$C_4^{(2)}$	$C_5^{(2)}$	$C_6^{(2)}$	$C_7^{(2)}$
$C_1^{(1)}$				1	0,55	0,98	0,02
$C_2^{(1)}$	0,13	0,43				0,02	0,98
$C_3^{(1)}$	0,87	0,57	1		0,45		

Tab.: $\Omega^{(2,1)}$

The intercluster similarity

- From these matrices ($\Omega^{(i,j)}$ and $\Omega^{(j,i)}$), a **similarity measure** is computed to compare two clusters of two different results.
- This **intercluster similarity** is evaluated by observing their intersection of the clusters and by taking into account the distribution of the cluster $C_k^{(i)}$ in all the clusters of $\mathcal{C}^{(j)}$ as follows :

$$S\left(C_k^{(i)}, \mathcal{C}_l^{(j)}\right) = \rho_k^{(i,j)} \alpha_{l,k}^{(j,i)} \tag{2}$$

where

$$\rho_k^{(i,j)} = \sum_{r=1}^{n_j} (\alpha_{k,r}^{(i,j)})^2 \tag{3}$$

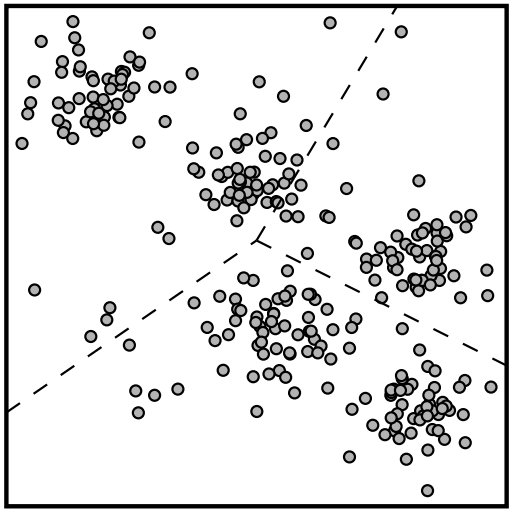
The corresponding cluster

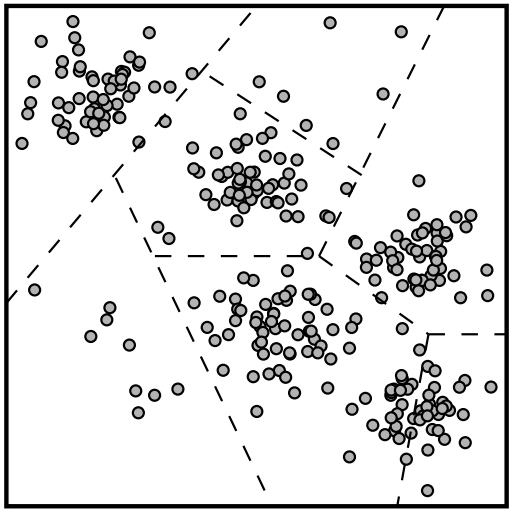
- Then, this similarity measure is used to compute the **corresponding cluster** of each cluster, which is the most similar cluster in an other result :

$$\psi\left(C_k^{(i)}, \mathcal{C}^{(j)}\right) = \arg \max_{C_l^{(j)} \in \mathcal{C}^{(j)}} S\left(C_k^{(i)}, C_l^{(j)}\right) \tag{4}$$

- This correspondence between the clusters of the different results is used to identify **the conflicts** between the different results.

The corresponding cluster





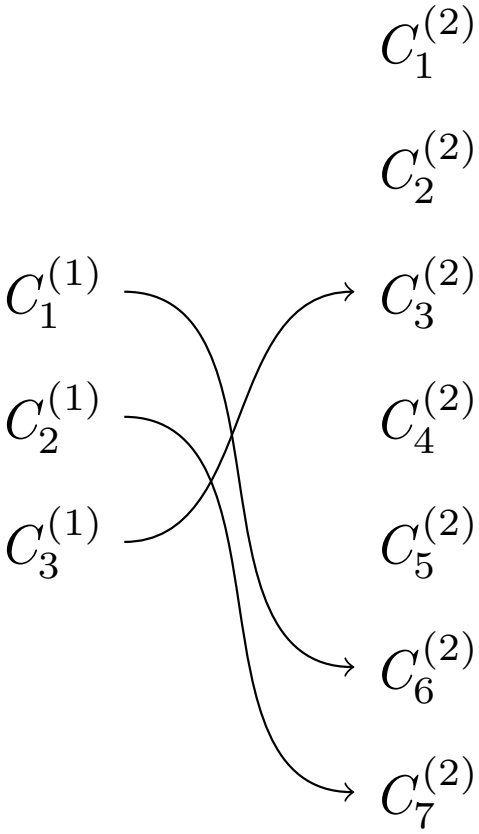


Fig.: $\psi\left(C_k^{(1)}, \mathcal{C}^{(2)}\right)$

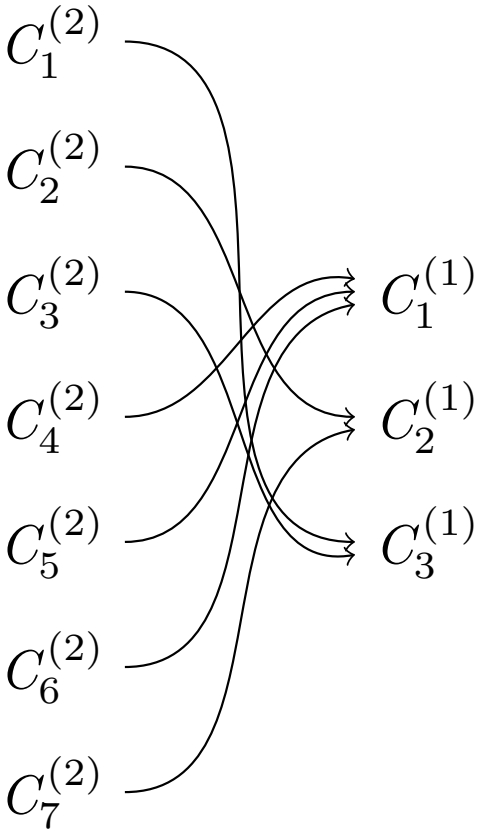


Fig.: $\psi\left(C_k^{(2)}, \mathcal{C}^{(1)}\right)$

The detection of the conflicts

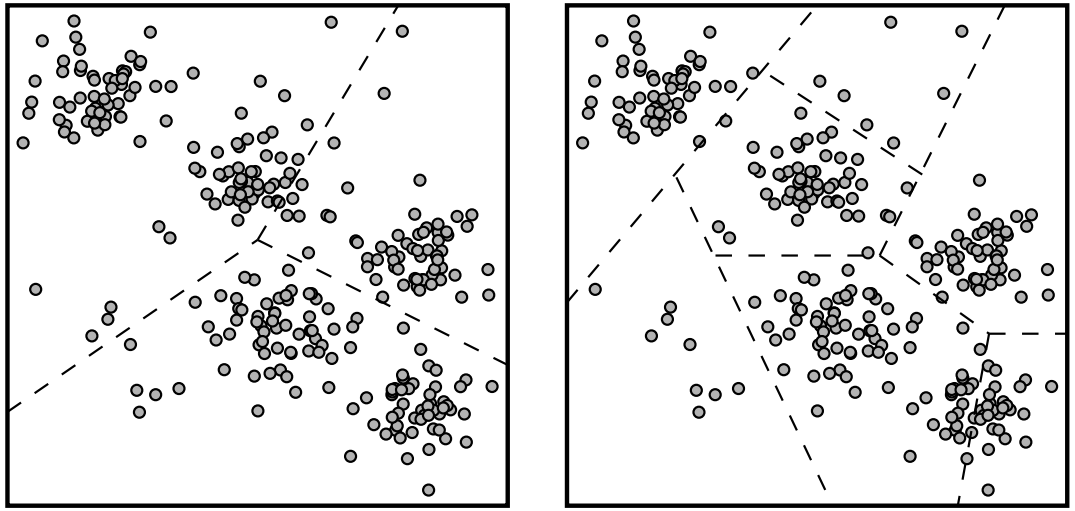
- The **detection of the conflicts** consists in seeking in $\mathbb{C}$ all the couples $\left(C_k^{(i)}, \mathcal{C}^{(j)}\right)$, $i \neq j$, such as $S\left(C_k^{(i)}, \psi\left(C_k^{(i)}, \mathcal{C}^{(j)}\right)\right) < 1$, which means that the cluster $C_k^{(i)}$ can not be exactly found in the result $\mathcal{C}^j$:

$$\texttt{conflicts}(\mathbb{C}) = \left\{ (C_k^{(i)}, \mathcal{C}^{(j)}) : i \neq j, S\left(C_k^{(i)}, \psi\left(C_k^{(i)}, \mathcal{C}^{(j)}\right)\right) < 1 \right\} \quad (5)$$

- Each conflict $\mathcal{K}_k^{(i,j)}$ is identified by **one cluster** $C_k^{(i)}$ and **one result** $\mathcal{C}^{(j)}$. Its importance, $CI\left(\mathcal{K}_k^{(i,j)}\right)$, is computed according to the intercluster similarity :

$$CI\left(\mathcal{K}_k^{(i,j)}\right) = 1 - S\left(C_k^{(i)}, \psi\left(C_k^{(i)}, \mathcal{C}^{(j)}\right)\right) \quad (6)$$

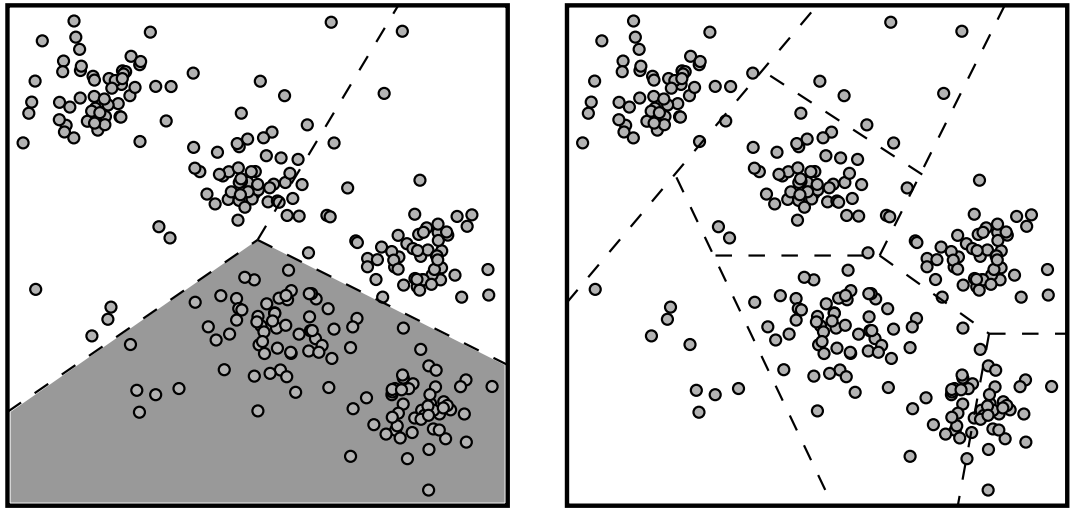
Example of a conflict



(a) 3 clusters (b) 7 clusters

Fig.: Example of conflict between two results.

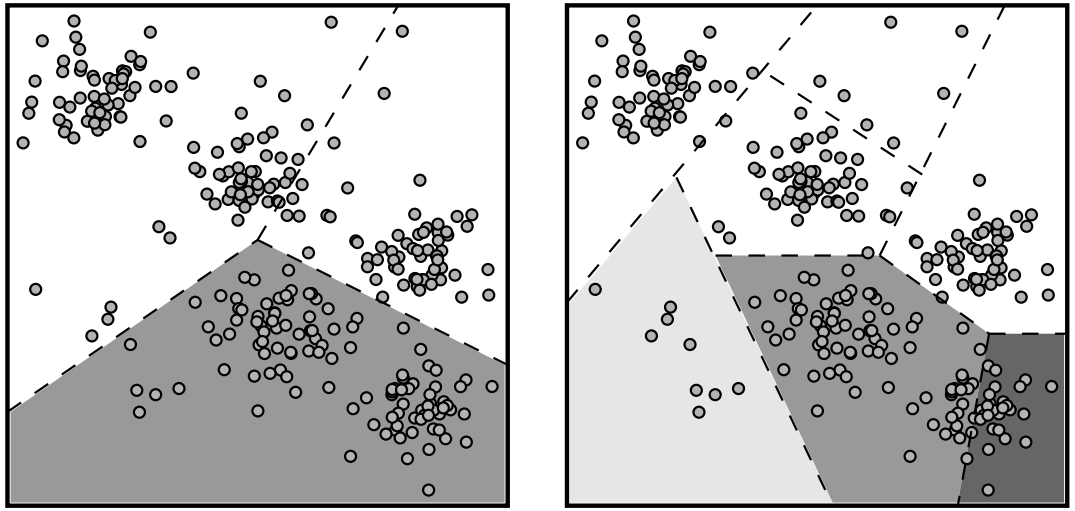
Example of a conflict



(a) 3 clusters (b) 7 clusters

Fig.: Example of conflict between two results.

Example of a conflict



(a) 3 clusters (b) 7 clusters

Fig.: Example of conflict between two results.

- $CI\left(\mathcal{K}_2^{(1,2)}\right) = 0.573$

The conflict resolution

- The local resolution of a conflict $\mathcal{K}_k^{(i,j)}$ consists in applying an operator on each result involved in the conflict, $\mathcal{C}^{(i)}$ and $\mathcal{C}^{(j)}$, to try to improve their similarity.
- The operators which can be applied to a result are the following :
  - **merging** of clusters : some clusters are merged together,
  - **splitting** of a cluster into subclusters : a clustering is applied to the objects of a cluster to produce subclusters,
  - **reclustering** of a group of objects : one cluster is removed and its objects are reclassified in all the other existing clusters.
- The operator to apply is chosen according to the number of clusters involved in the conflict.

The similarity evaluation

- However, the application of the two operators (each one on a different result) **is not always relevant**.
- To evaluate the similarity between two results, we defined a criterion γ, called **local similarity criterion**.
- It is based on the intercluster similarity S and a quality criterion δ.

$$\gamma^{(i,j)} = \frac{1}{2}\left(p_s.\left(\frac{1}{n_i}\sum_{k=1}^{n_i}\omega_k^{(i,j)} + \frac{1}{n_j}\sum_{k=1}^{n_j}\omega_k^{(j,i)}\right) + p_q.\left(\delta^{(i)} + \delta^{(j)}\right)\right) \quad (7)$$

where

$$\omega_k^{(i,j)} = S\left(C_k^{(i)}, \psi\left(C_k^{(i)}, \mathcal{C}^{(j)}\right)\right) \quad (8)$$

and, p_q and p_s are given by the user ($p_q + p_s = 1$).

The global agreement coefficient

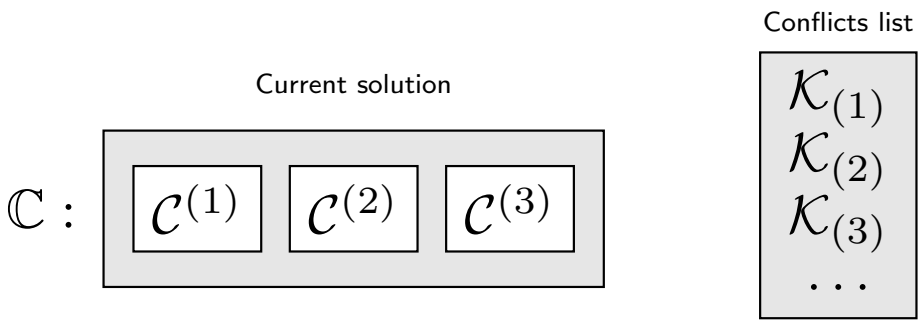
- After the resolutions of the local conflicts, a **global application** of the modifications proposed by the refinement step is decided if their application improve the quality of the global result.
- The **global agreement coefficient $\mathbf{\Gamma}$** is evaluated according to all the local similarities between each couple of results as follows :

$$\mathbf{\Gamma} = \frac{1}{m} \sum_{i=1}^{m} \Gamma^i \tag{9}$$

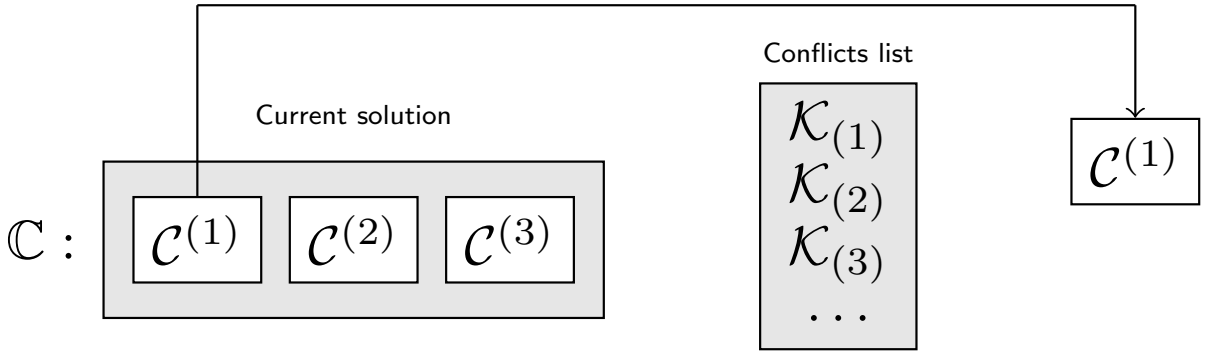
where

$$\Gamma^i = \frac{1}{m-1} \sum_{\substack{j=1 \\ j \neq i}}^{m} \gamma^{i,j} \tag{10}$$

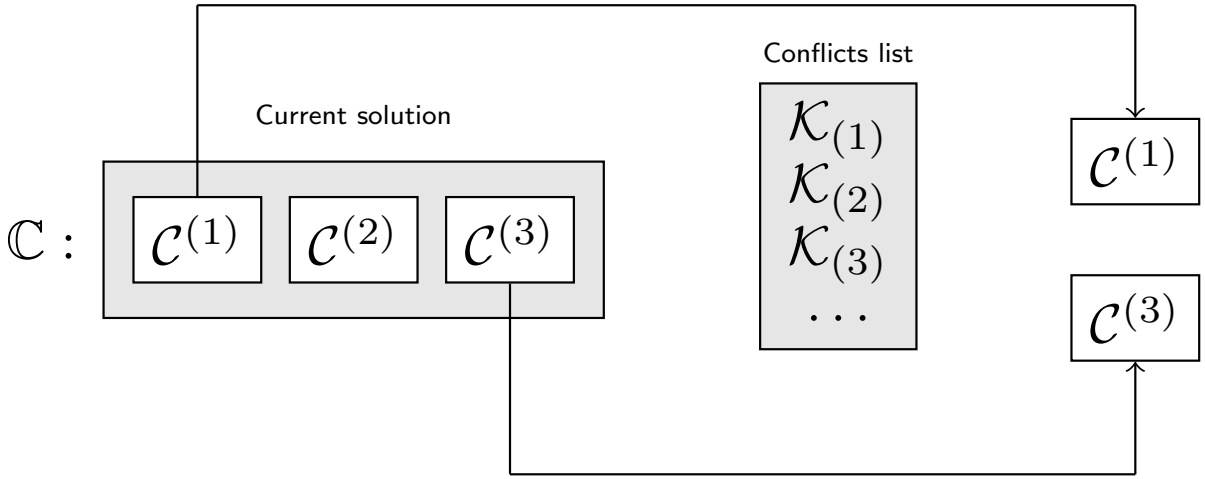
Workflow of the approach



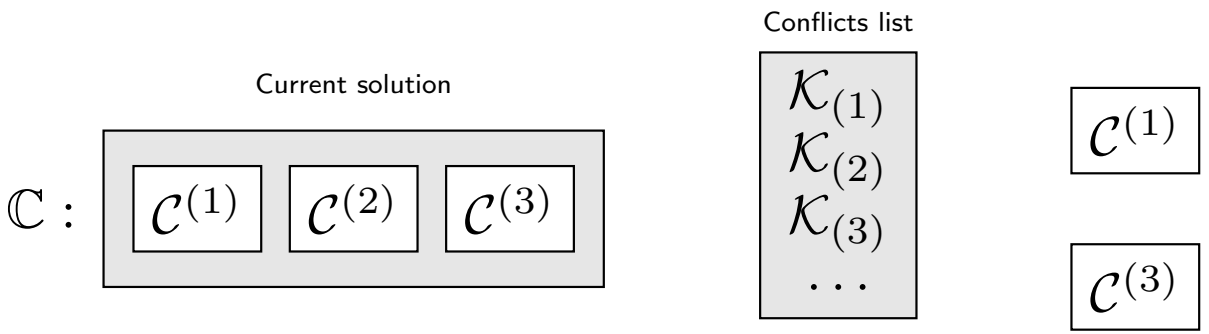
Workflow of the approach



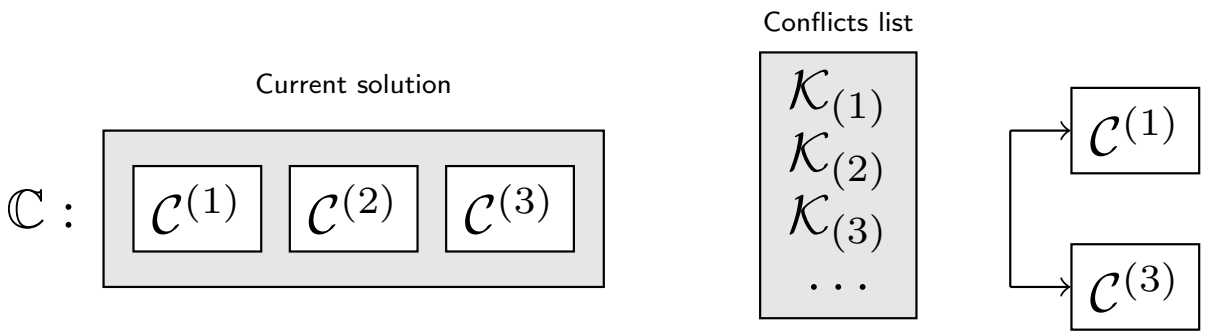
Workflow of the approach



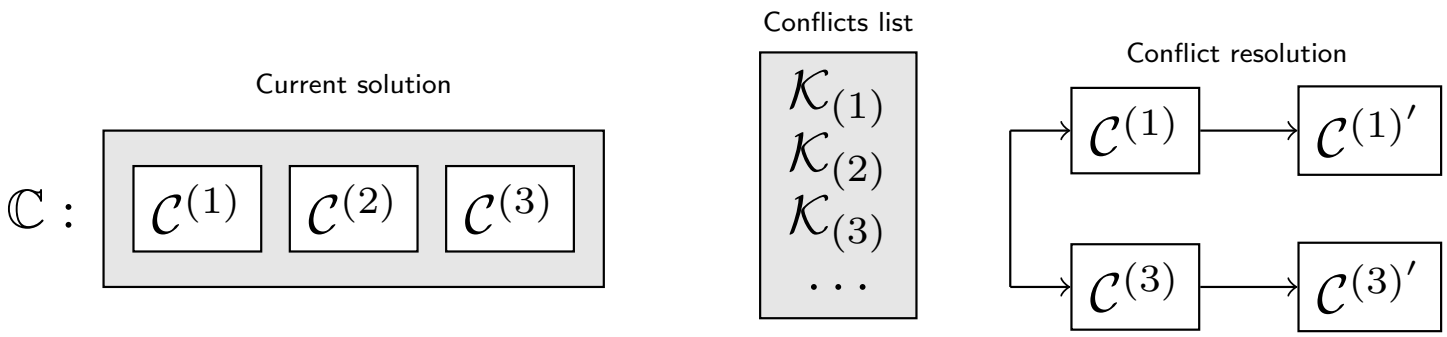
Workflow of the approach



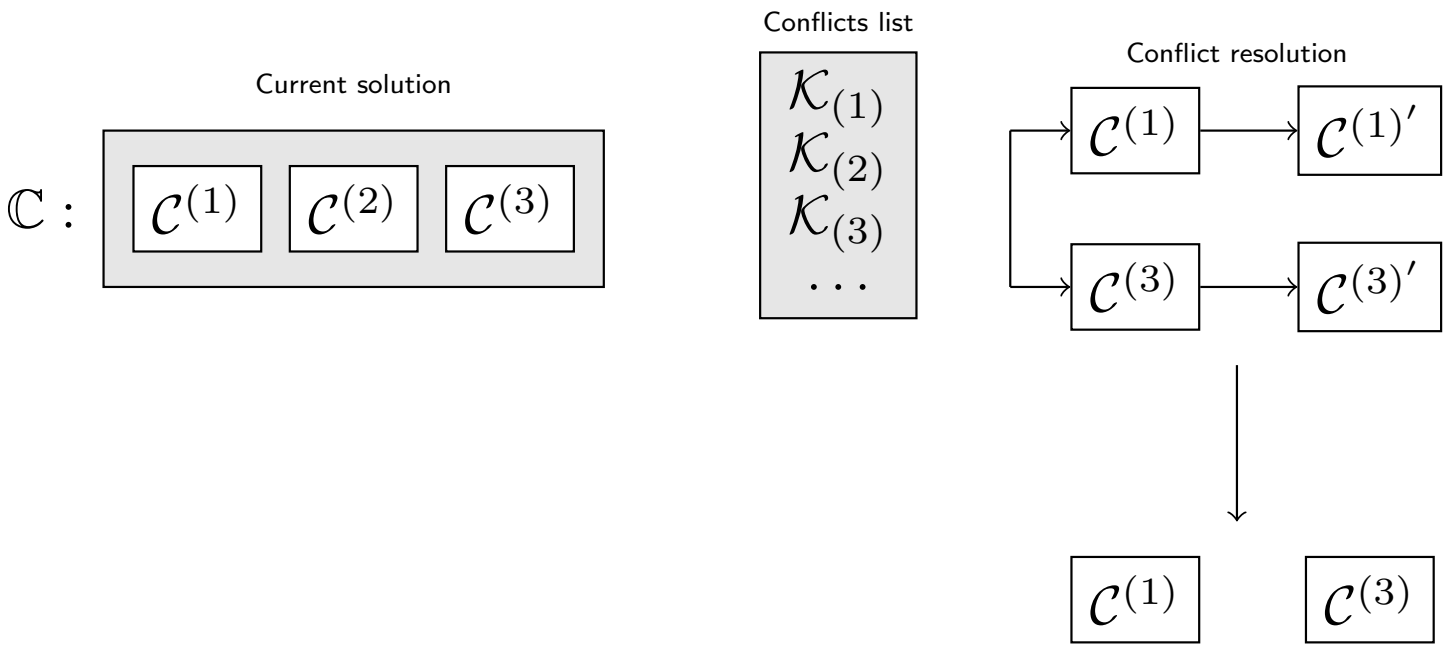
Workflow of the approach



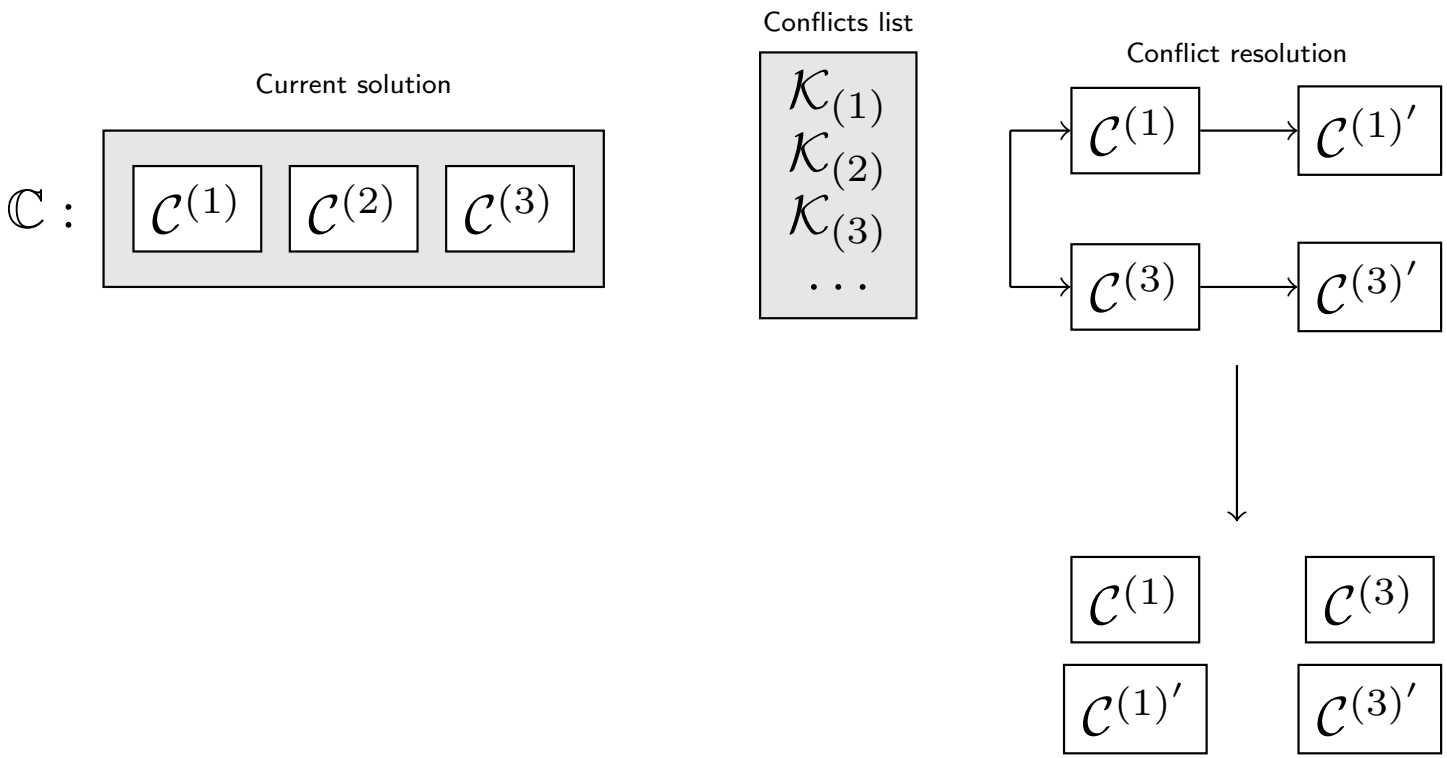
Workflow of the approach



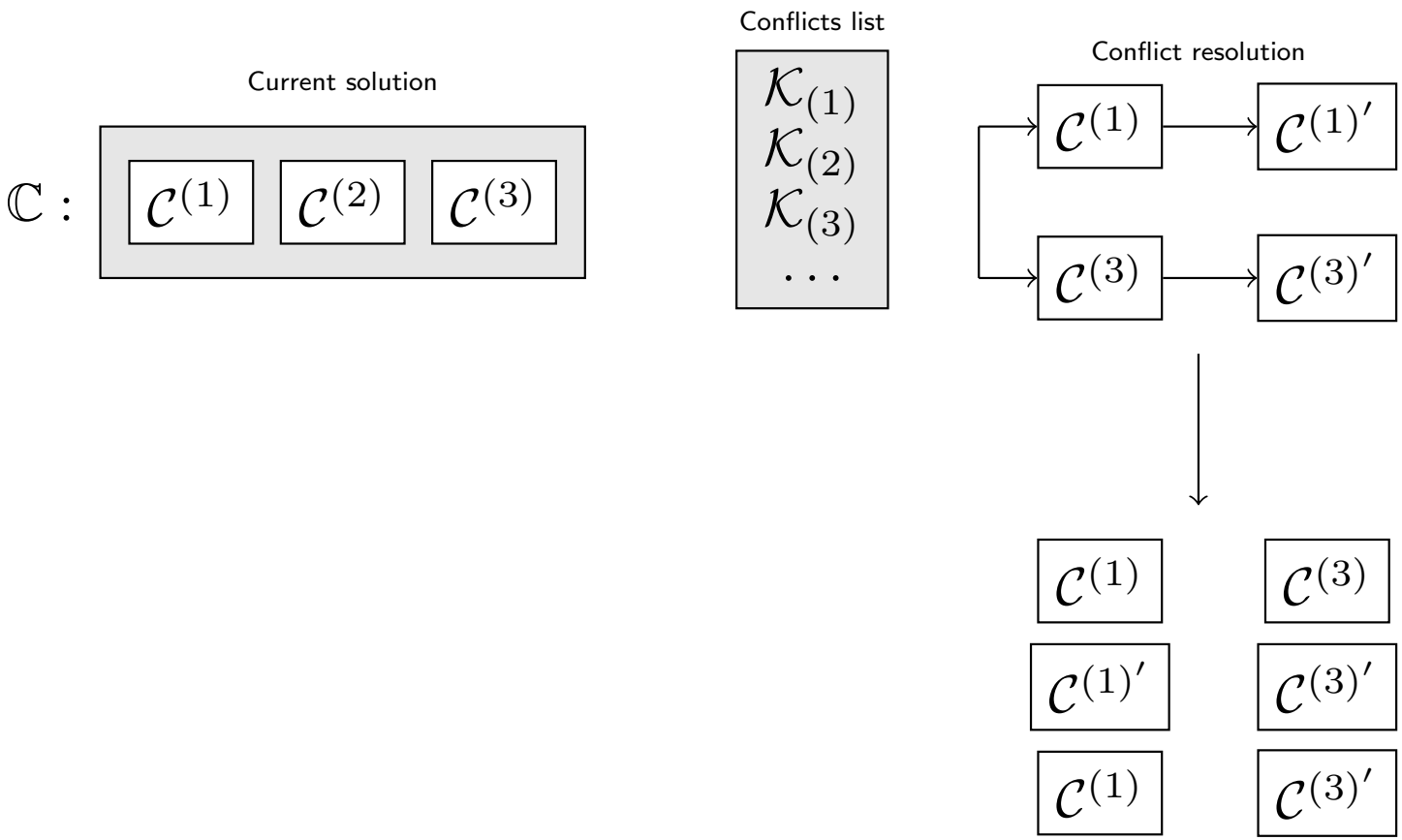
Workflow of the approach



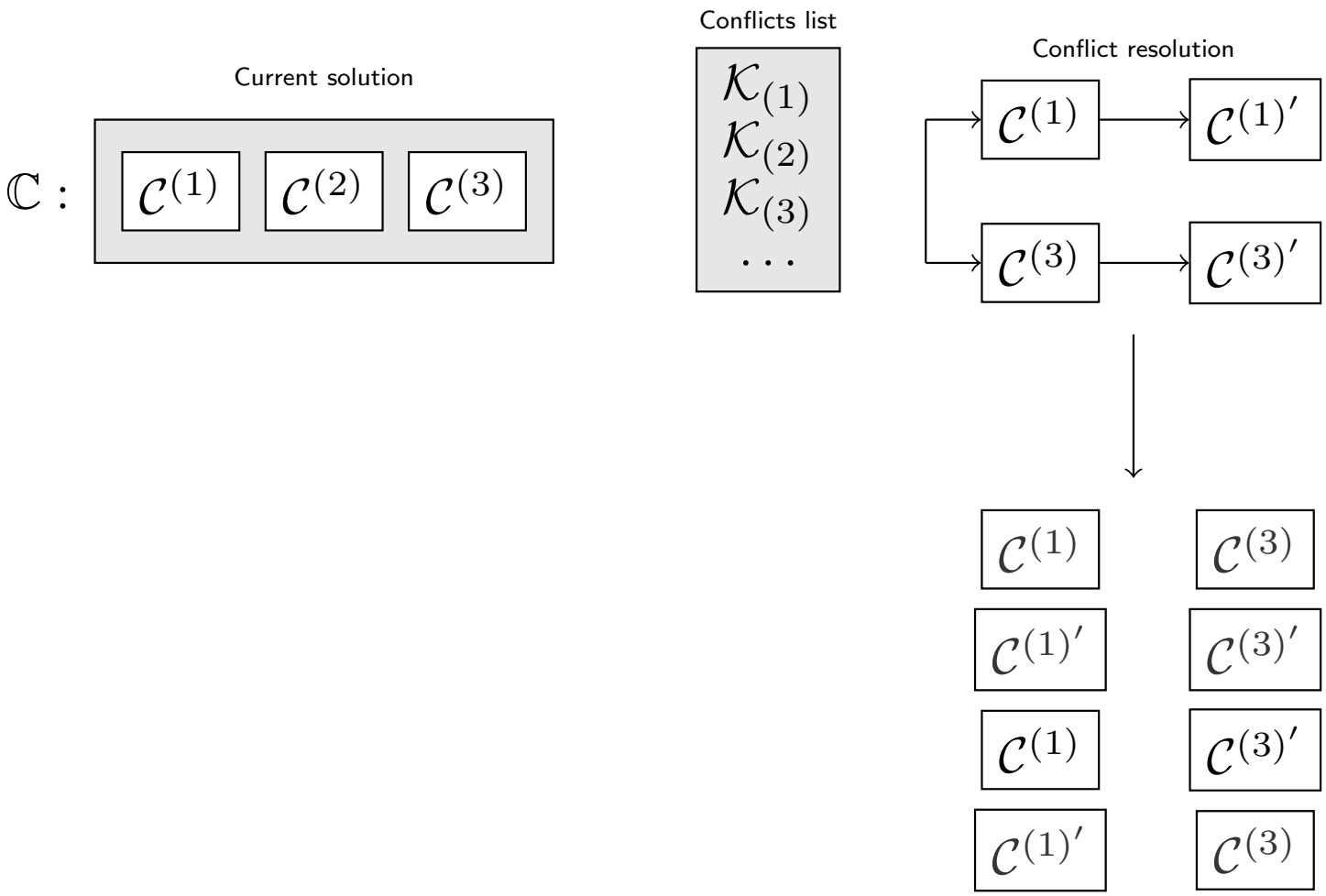
Workflow of the approach



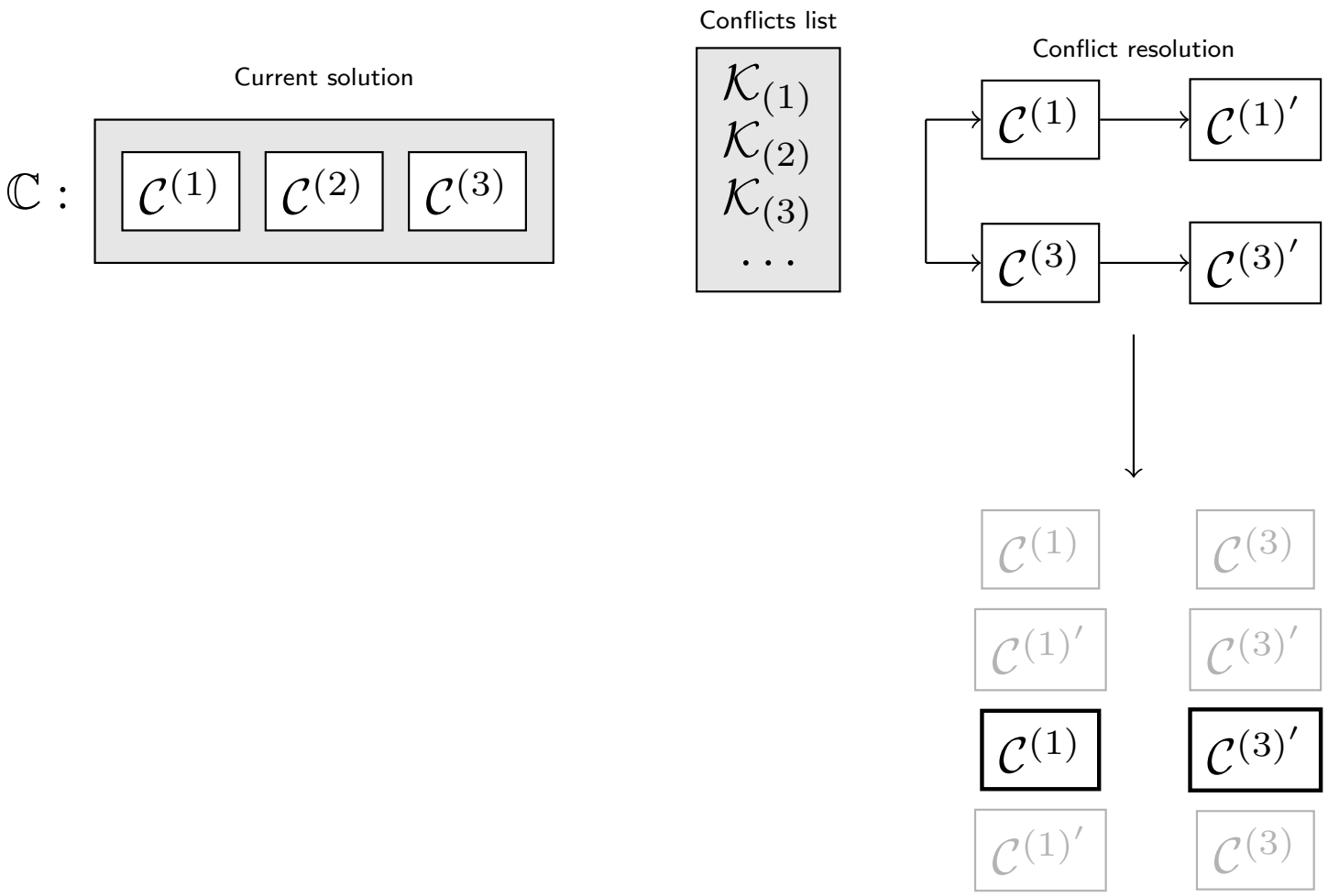
Workflow of the approach



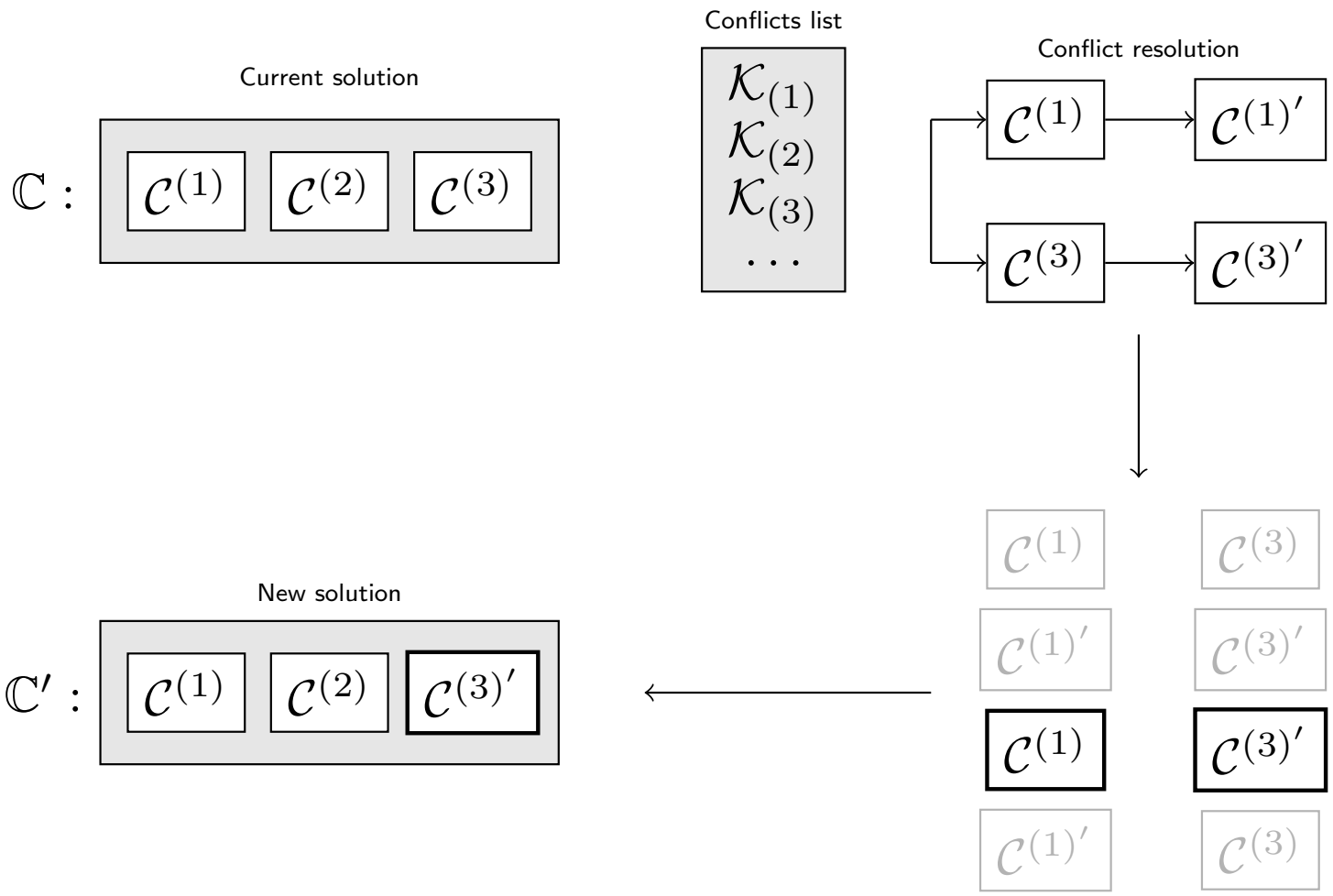
Workflow of the approach



Workflow of the approach



Workflow of the approach



Worst conflict choice (WCC)

- This first approach consists in choosing the worst conflict, i.e. the one having the highest conflict importance.

$$\mathcal{K} := \arg \max_{\mathcal{K}_{(i)} \in \breve{\mathcal{K}}} CI\left(\mathcal{K}_{(i)}\right) \tag{11}$$

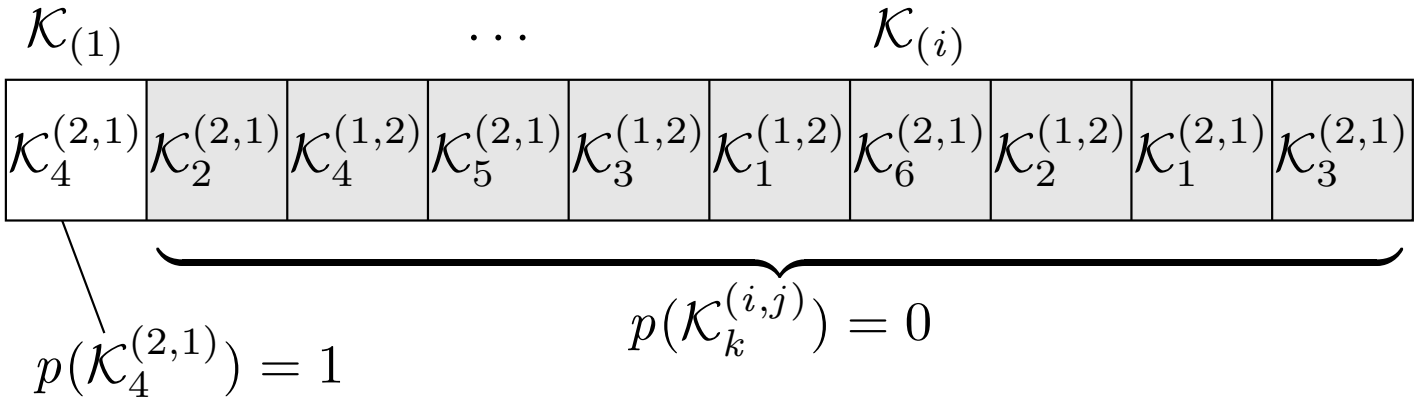


Fig.: WCC

Stochastic conflict choice (SCC)

- This approach consist in choosing randomly a conflict to solve in the list of conflicts. This naive strategy assumes that solving a conflict with a high importance is not always relevant.

$$\mathcal{K} := \texttt{random}\left(\breve{\mathcal{K}}\right) \tag{12}$$

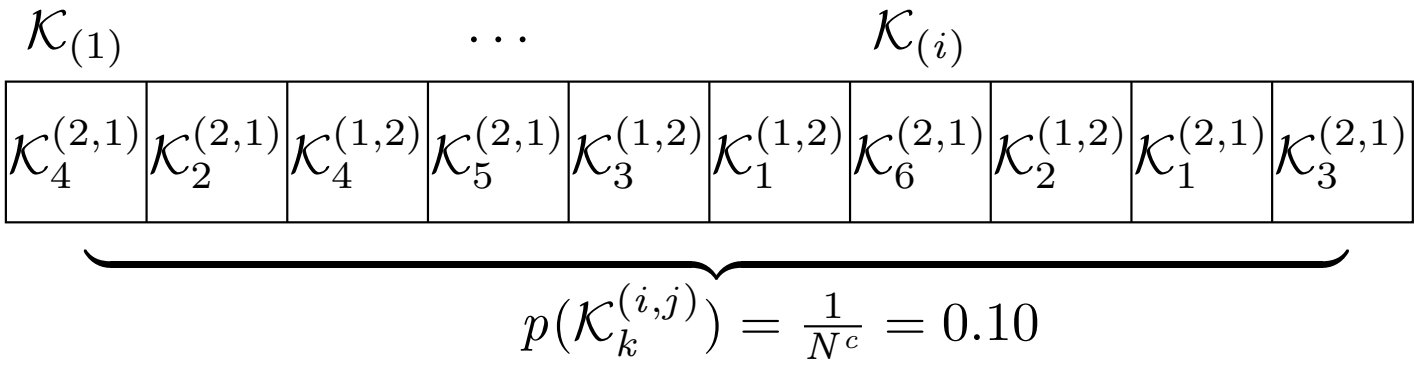


Fig.: SCC

Roulette-wheel conflict choice (R-WCC)

- This approach is based on roulette-wheel selection or fitness proportionate selection well known in the field of evolutionary optimization. The conflict importance is used to associate a probability of selection with each conflict. In this case, the probability of each conflict is defined as $p(\mathcal{K}_{(i)}) = \frac{CI\left(\mathcal{K}_{(i)}\right)}{\sum_{j=1}^{N_c} CI\left(\mathcal{K}_{(j)}\right)}$ with N_c the number of conflicts.

$$\mathcal{K} := \mathcal{K}_{(i)} \Big| \left(\sum_{j=0}^{i-1} p(\mathcal{K}_{(j)}) \leq \nu \wedge \sum_{k=0}^{i+1} p(\mathcal{K}_{(k)}) > \nu \right) \tag{13}$$

$\mathcal{K}_{(1)}$		$\cdots$		$\mathcal{K}_{(i)}$				
$\mathcal{K}_4^{(2,1)}$	$\mathcal{K}_2^{(2,1)}$	$\mathcal{K}_4^{(1,2)}$	$\mathcal{K}_5^{(2,1)}$	$\mathcal{K}_3^{(1,2)}$	$\mathcal{K}_1^{(1,2)}$	$\mathcal{K}_6^{(2,1)}$	$\cdots$	$\cdots$

$$p\left(\mathcal{K}_4^{(2,1)}\right) = \frac{CI(\mathcal{K}_4^{(2,1)})}{\sum_{j=1}^{N^c} CI(\mathcal{K}_{(j)})} = 0.17$$

The global agreement coefficient

- **Evolutionary algorithm** (EA) is a generic population-based metaheuristic optimization algorithm.
- An EA uses some mechanisms inspired by **biological evolution** : mutation, recombination, and selection.
- Candidate solutions to the optimization problem play the role of **individuals** in a **population**.
- A **fitness function** evaluates the quality of the solutions.
- An **evolution** of the population takes place after by repeated application of the operators (mutation, recombination, and selection).

The global agreement coefficient

- In our framework, a **solution** will be a set of clustering results : $\mathbb{C}$
- The **fitness** function will be the global agreement value of the set : $\mathbf{\Gamma}(\mathbb{C})$
- Instead of modifiying iteratively one solution, we will have a **population** of solutions evolving together by **sharing information**

$$\mathbb{C} : \boxed{\boxed{\mathcal{C}^{(1)}}\ \boxed{\mathcal{C}^{(2)}}\ \boxed{\cdots}\ \boxed{\mathcal{C}^{(n)}}}$$

Fig.: Representation of one solution.

The evolutionary steps

- **Initialisation of the population** : random application of operators (split, merge, reclustering) on the original solution
- **Selection** : tournament selection based on the global coefficient $\mathbf{\Gamma}(\mathbb{C})$
- **Crossover** : creation of a new solution by mixing two existing ones
- **Mutation** : resolution of a randomly selected conflict

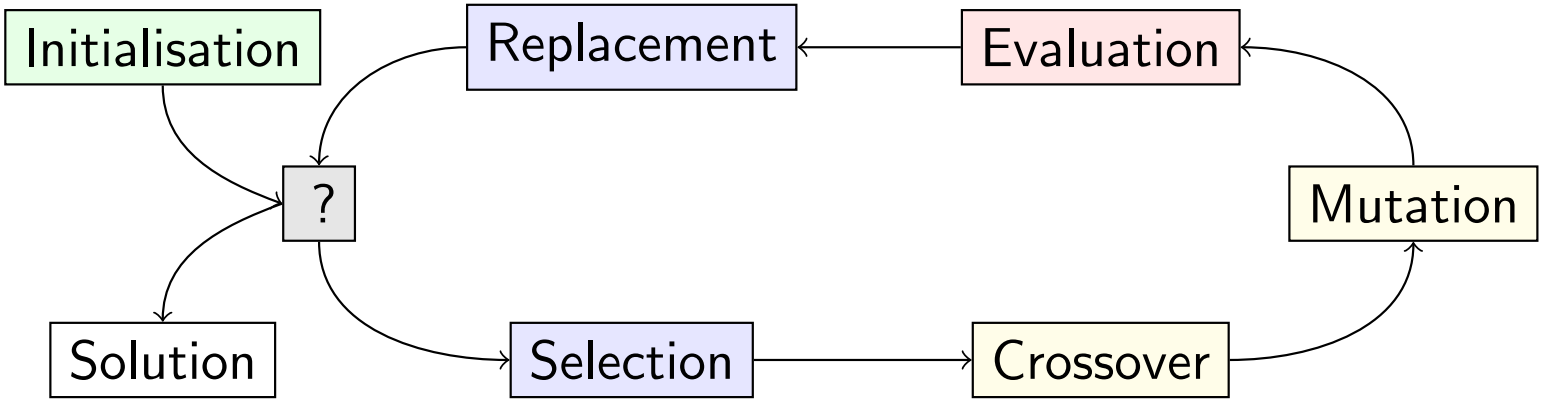


Fig.: The evolutionary process.

Workflow of the approach

Initial solution

$\mathbb{C}$: $\boxed{\boxed{\mathcal{C}^{(1)}}\ \boxed{\mathcal{C}^{(2)}}\ \boxed{\mathcal{C}^{(3)}}}$

Workflow of the approach

Initial solution

$\mathbb{C}$: $\mathcal{C}^{(1)}$ $\mathcal{C}^{(2)}$ $\mathcal{C}^{(3)}$

$\mathbb{C}'$: $\mathcal{C}^{(1)}$ $\mathcal{C}^{(2)'}$ $\mathcal{C}^{(3)'}$

Workflow of the approach

Initial solution

$\mathbb{C}$: $\mathcal{C}^{(1)}$ $\mathcal{C}^{(2)}$ $\mathcal{C}^{(3)}$

$\mathbb{C}'$: $\mathcal{C}^{(1)}$ $\mathcal{C}^{(2)'}$ $\mathcal{C}^{(3)'}$

$\mathbb{C}''$: $\mathcal{C}^{(1)''}$ $\mathcal{C}^{(2)}$ $\mathcal{C}^{(3)''}$

Workflow of the approach

Initial solution

$\mathbb{C}$: $\boxed{\mathcal{C}^{(1)}}\ \boxed{\mathcal{C}^{(2)}}\ \boxed{\mathcal{C}^{(3)}}$

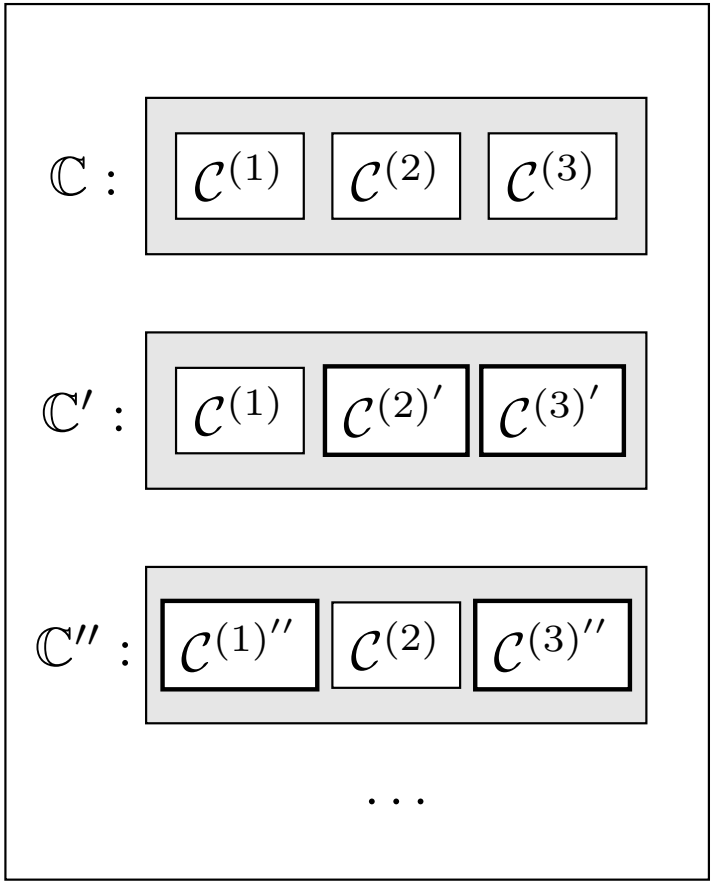
$\mathbb{C}'$: $\boxed{\mathcal{C}^{(1)}}\ \boxed{\mathcal{C}^{(2)'}}\ \boxed{\mathcal{C}^{(3)'}}$

$\mathbb{C}''$: $\boxed{\mathcal{C}^{(1)''}}\ \boxed{\mathcal{C}^{(2)}}\ \boxed{\mathcal{C}^{(3)''}}$

...

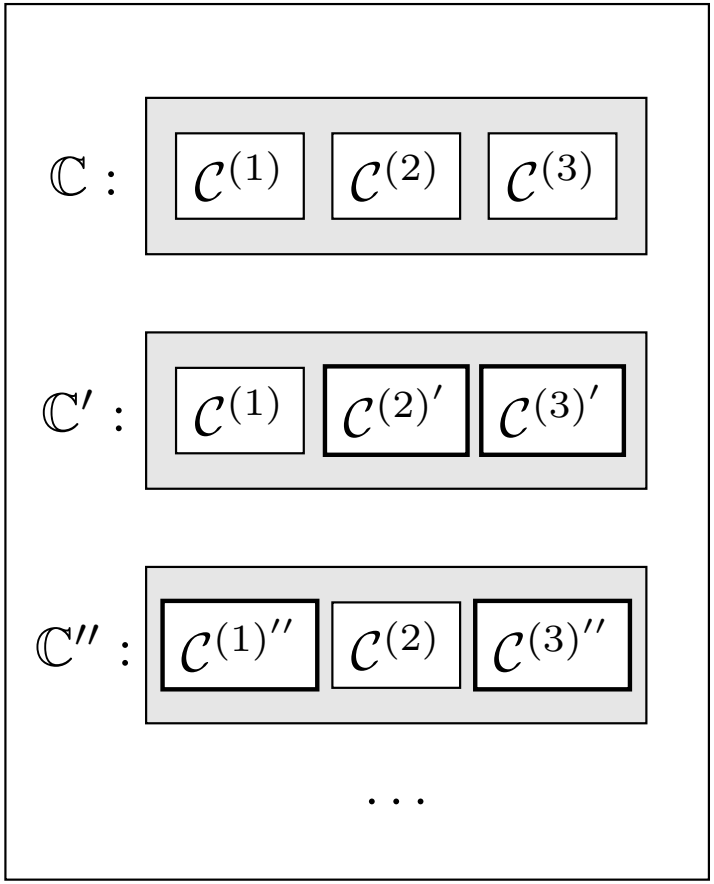
Workflow of the approach

Initial population

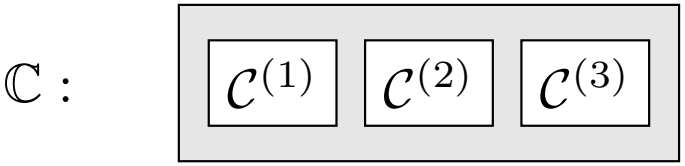


Workflow of the approach

Initial population

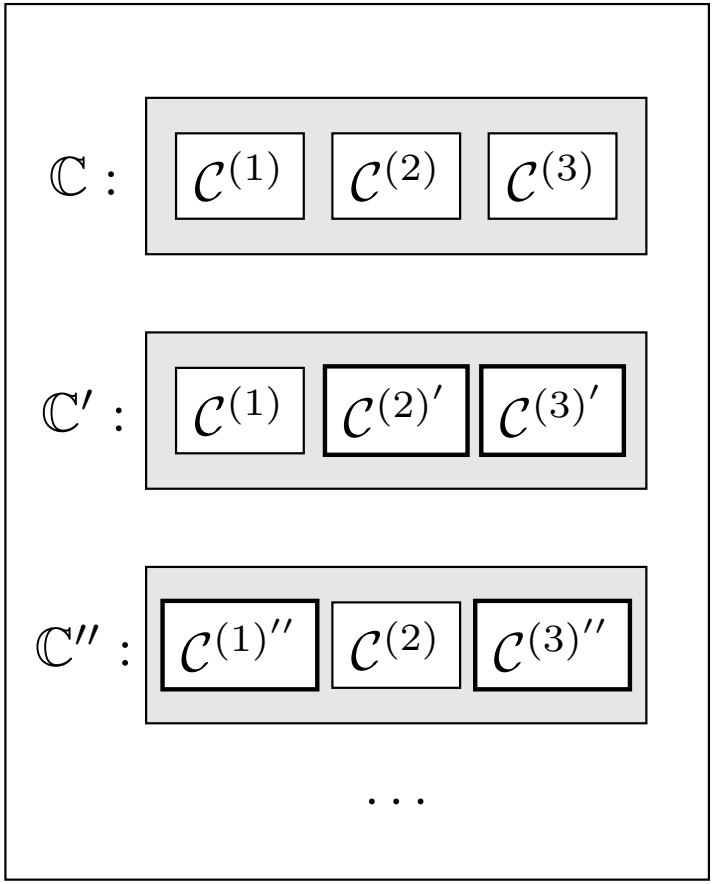


The cross over operator

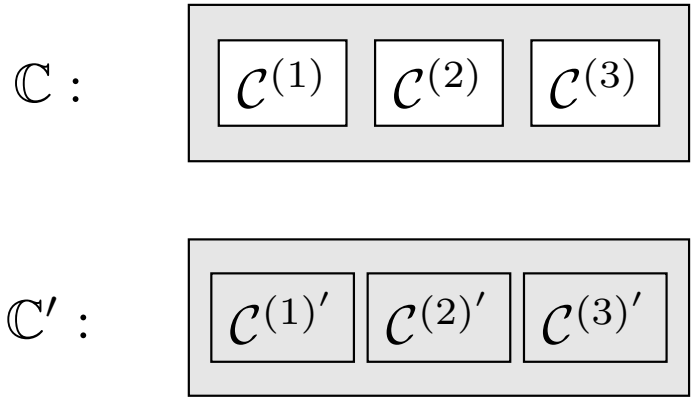


Workflow of the approach

Initial population

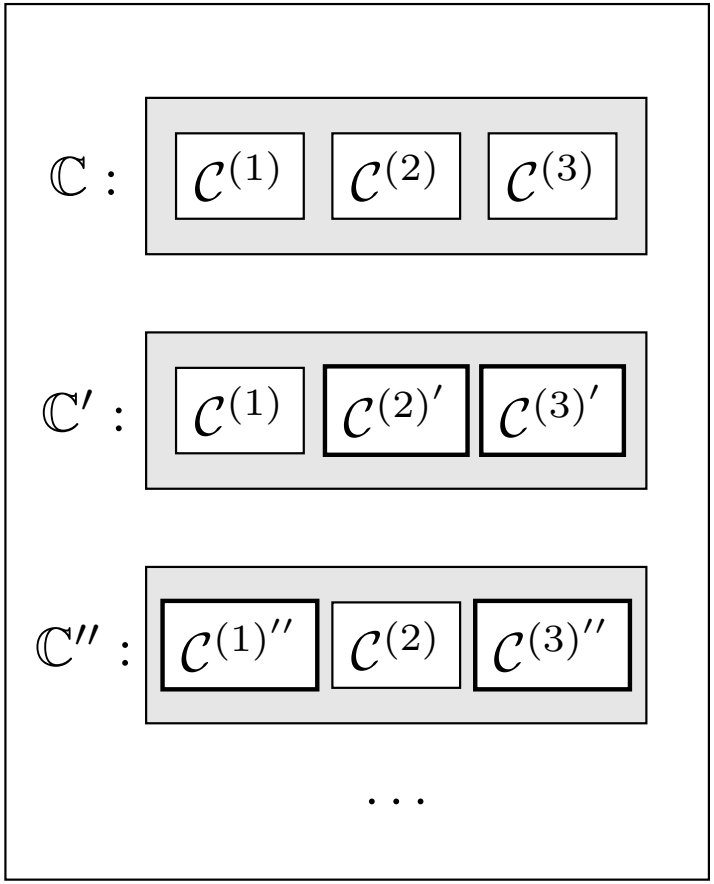


The cross over operator

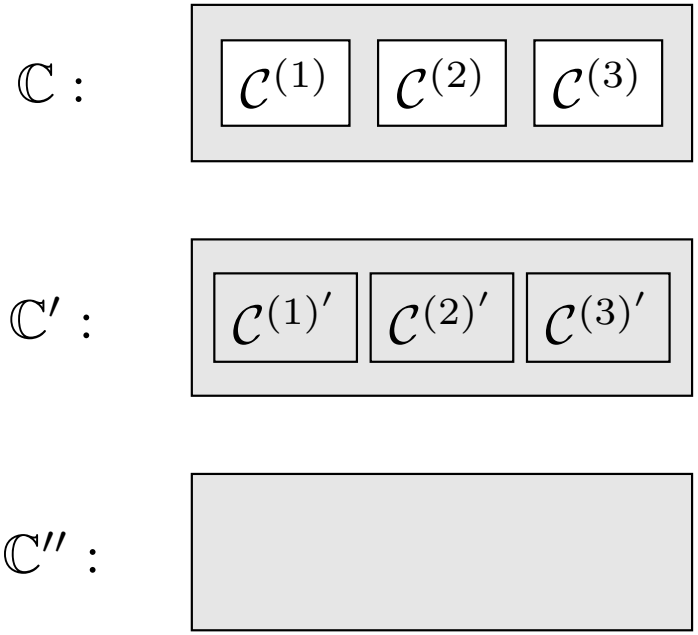


Workflow of the approach

Initial population

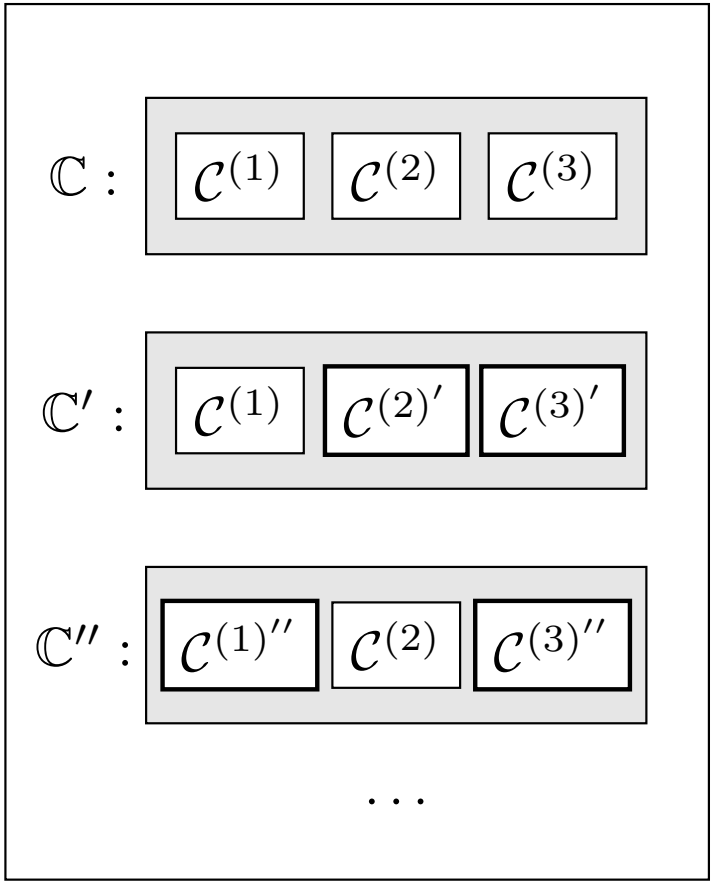


The cross over operator

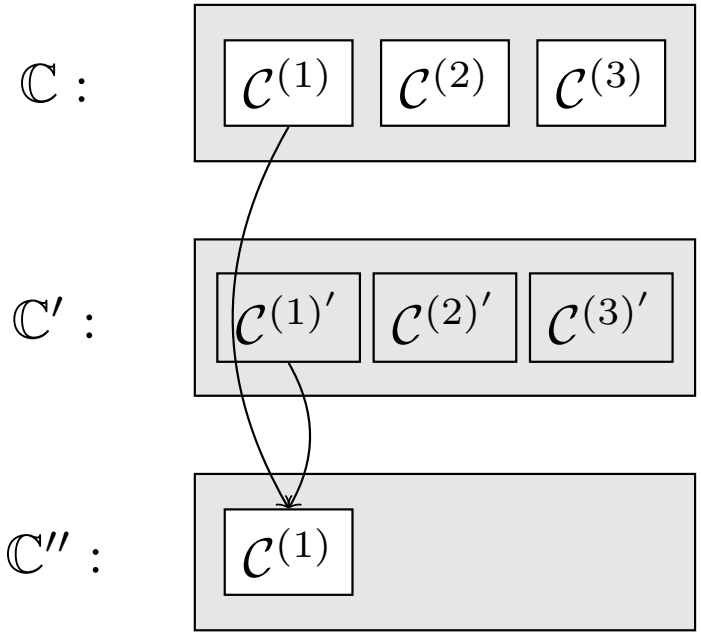


Workflow of the approach

Initial population

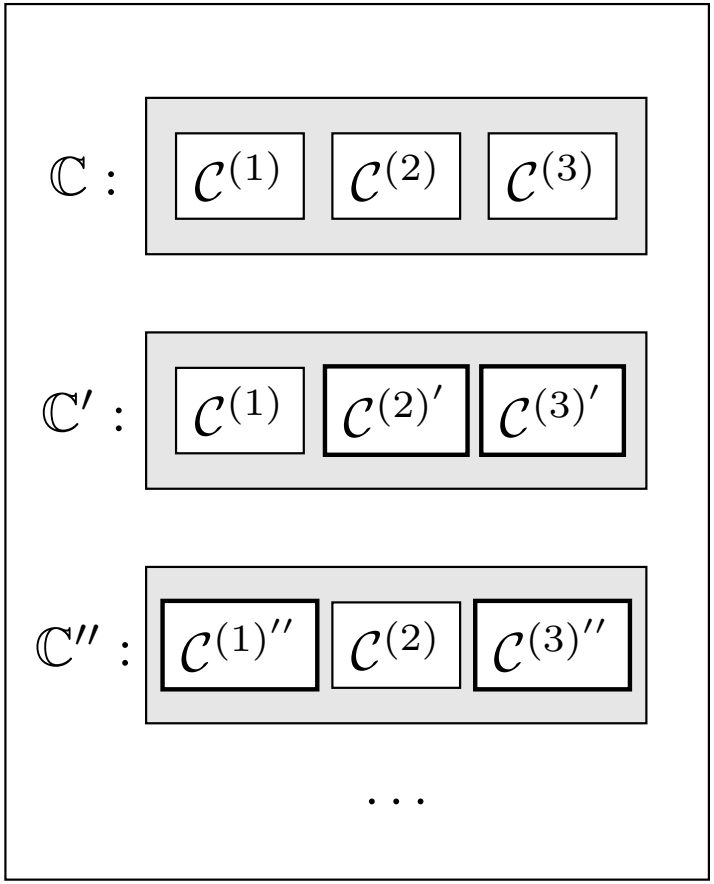


The cross over operator

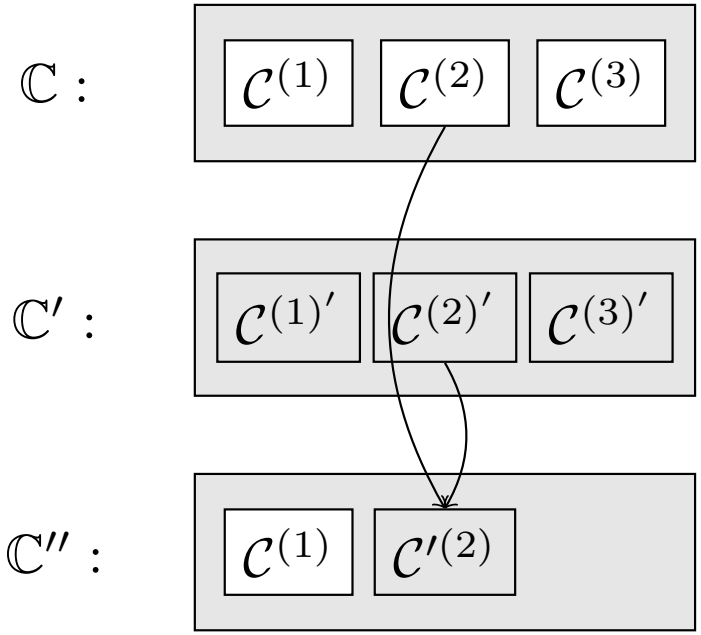


Workflow of the approach

Initial population

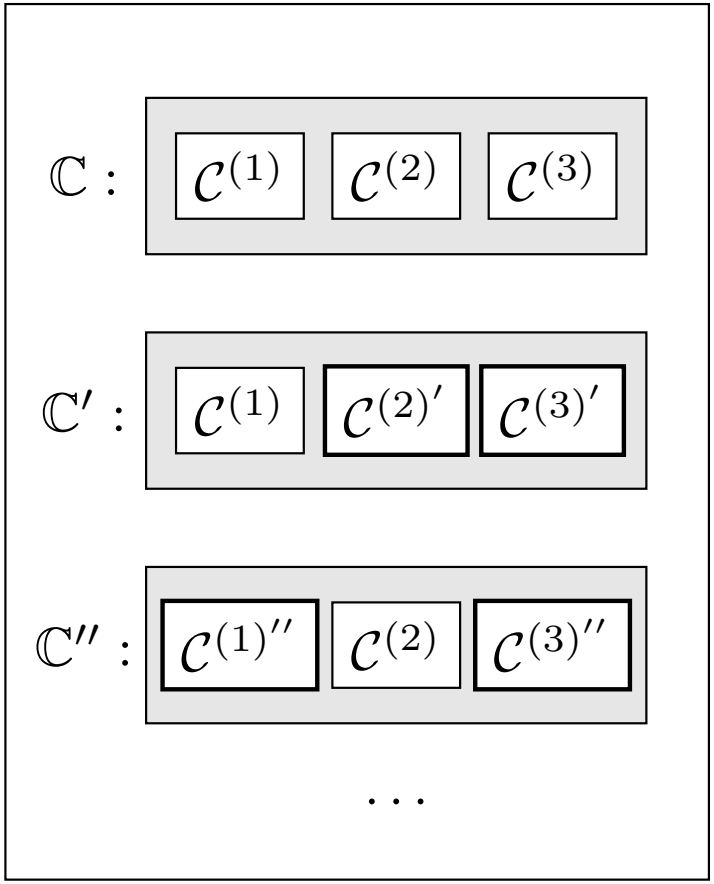


The cross over operator

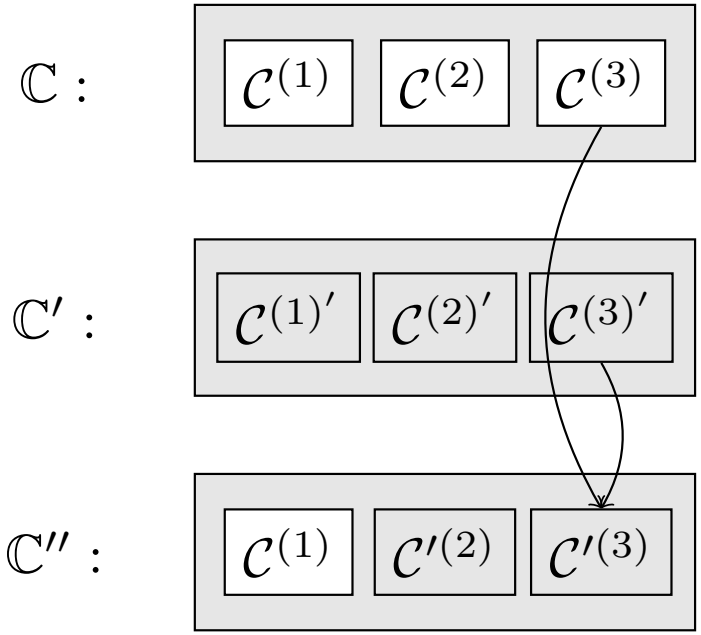


Workflow of the approach

Initial population

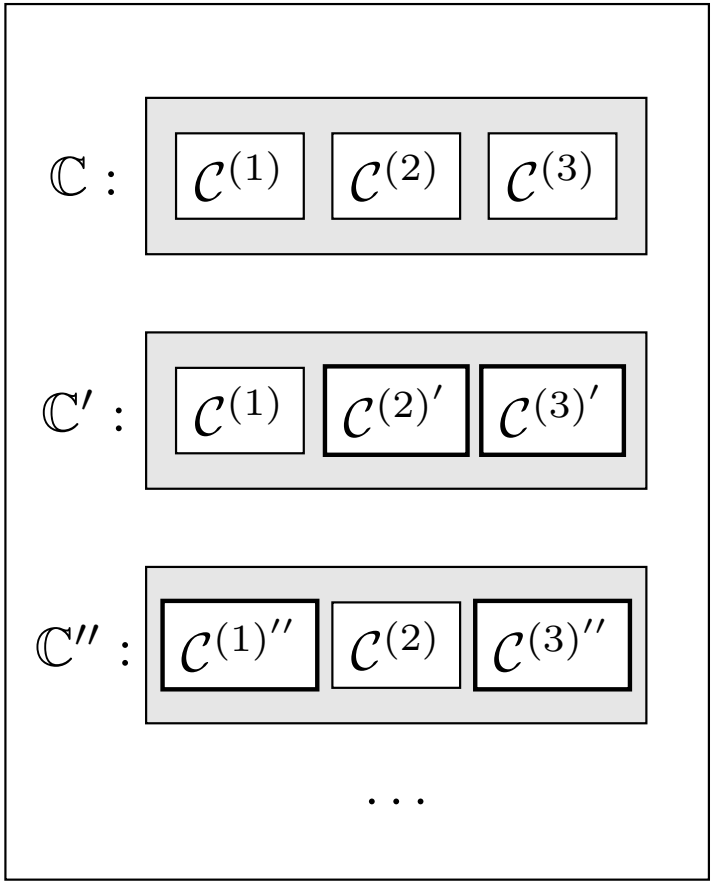


The cross over operator

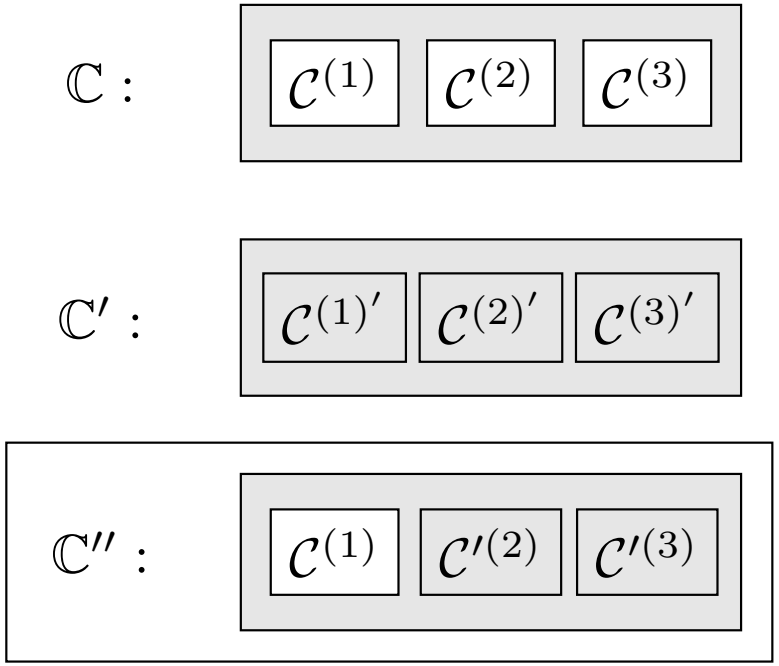


Workflow of the approach

Initial population

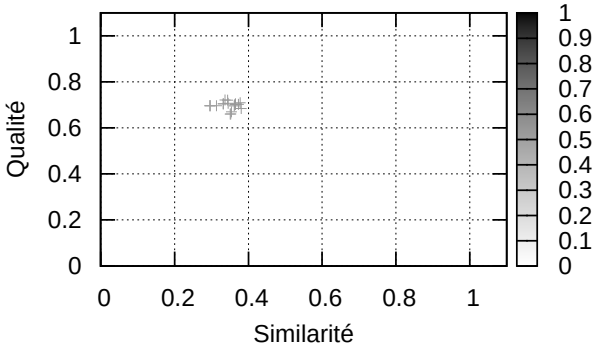


The cross over operator

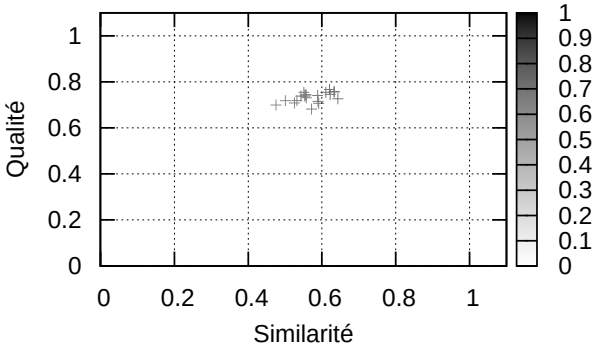


New solution

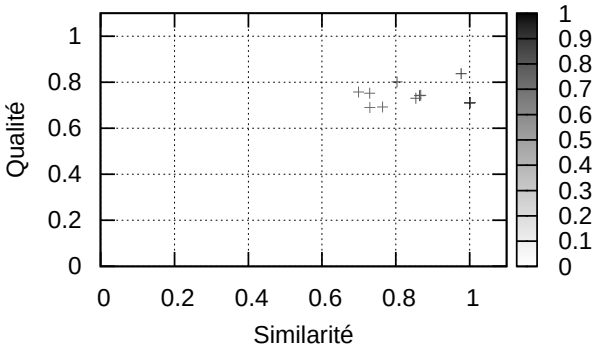
Example of the evolution of the population



(a) Initial population

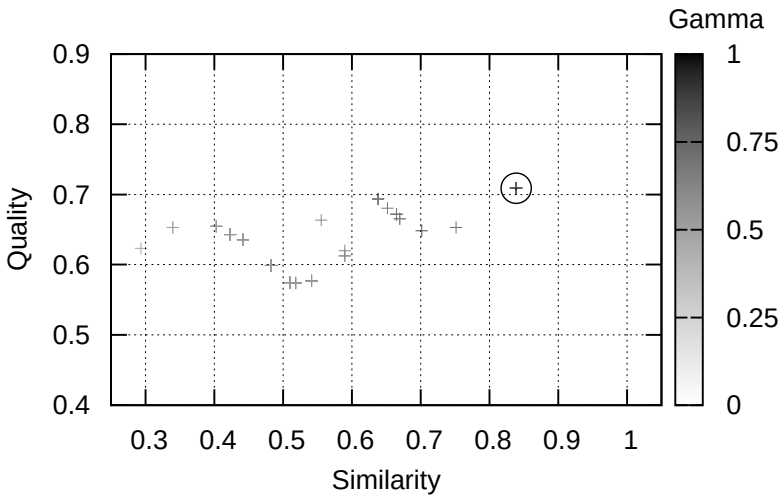


(b) Generation 50

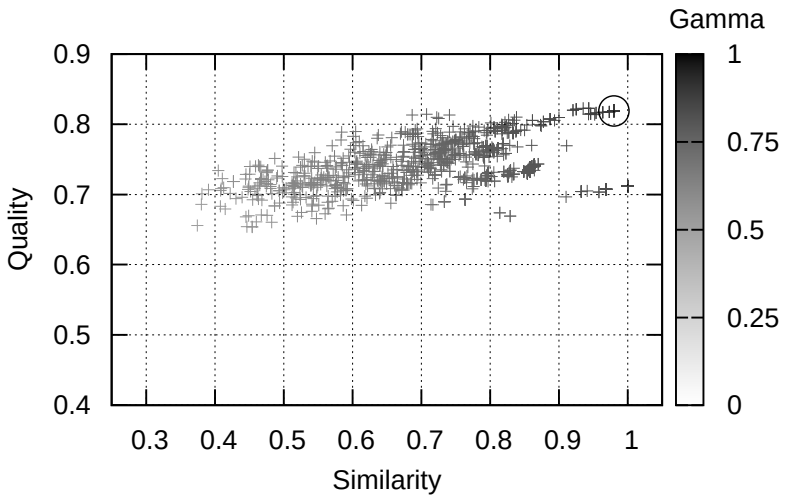


(c) Generation 100

Comparison of the approaches



(d) Solutions of the iterative approach



(e) Solutions of the genetic approach

Experiments

- We **evaluated** the four proposed approaches on synthetic datasets included in the Cluster generators package [a], and real-life datasets from the UCI repository.
- In all of the experiments, we used the **KMeans algorithm** as the base clustering method.
- Five instances of KMeans were randomly initialized with a number of clusters randomly picked in $[2; 10]$ to create the initial set of results.
- Then, **the different strategies of conflict resolution** were applied, each one individually, but starting from the same initial set of results (i.e. the same initial solution).

[a] http://dbkgroup.org/handl/generators/

Experiments

	Strategy	Γ	Rand	Jaccard	Folks & Mallows	F-Measure
2d-4c-no0	WCC	0.924	0.772 (±0.074)	0.872 (±0.045)	0.869 (±0.049)	0.824 (±0.066)
-	RCC	0.872	0.745 (±0.095)	0.857 (±0.057)	0.850 (±0.065)	0.808 (±0.080)
-	R-WCC	0.896	0.767 (±0.064)	0.871 (±0.037)	0.866 (±0.041)	0.831 (±0.045)
-	GR	**0.957**	**0.850 (±0.037)**	**0.919 (±0.021)**	**0.919 (±0.021)**	**0.890 (±0.028)**
2d-4c-no1	WCC	0.897	0.643 (±0.163)	0.788 (±0.110)	0.769 (±0.137)	0.661 (±0.247)
-	RCC	0.813	0.616 (±0.097)	0.768 (±0.066)	0.756 (±0.081)	0.693 (±0.074)
-	R-WCC	0.800	0.612 (±0.110)	0.765 (±0.074)	0.750 (±0.088)	0.692 (±0.086)
-	GR	**0.941**	**0.769 (±0.056)**	**0.871 (±0.039)**	**0.868 (±0.038)**	**0.817 (±0.052)**
10d-4c-no0	WCC	0.852	0.900 (±0.100)	0.947 (±0.055)	0.944 (±0.060)	0.913 (±0.093)
-	RCC	0.781	0.828 (±0.177)	0.898 (±0.111)	0.888 (±0.123)	0.856 (±0.148)
-	R-WCC	0.875	0.937 (±0.080)	0.967 (±0.044)	0.965 (±0.048)	0.946 (±0.073)
-	GR	**0.887**	**0.958 (±0.011)**	**0.979 (±0.006)**	**0.979 (±0.006)**	**0.967 (±0.009)**
10d-4c-no1	WCC	0.836	0.814 (±0.226)	0.890 (±0.144)	0.876 (±0.171)	0.806 (±0.296)
-	RCC	0.835	0.925 (±0.104)	0.958 (±0.062)	0.956 (±0.066)	0.939 (±0.091)
-	R-WCC	0.849	0.937 (±0.079)	0.967 (±0.043)	0.965 (±0.045)	0.954 (±0.058)
-	GR	**0.865**	**0.972 (±0.049)**	**0.985 (±0.026)**	**0.985 (±0.027)**	**0.979 (±0.037)**
iris	WCC	0.870	0.586 (±0.006)	0.763 (±0.006)	0.739 (±0.005)	0.615 (±0.010)
-	RCC	0.811	0.589 (±0.043)	0.759 (±0.029)	0.740 (±0.035)	0.634 (±0.027)
-	R-WCC	0.828	0.594 (±0.036)	0.762 (±0.024)	0.744 (±0.029)	0.636 (±0.027)
-	GR	**0.898**	**0.608 (±0.020)**	**0.771 (±0.009)**	**0.756 (±0.015)**	**0.638 (±0.018)**
wine	WCC	0.740	0.712 (±0.099)	0.831 (±0.065)	0.826 (±0.074)	0.734 (±0.120)
-	RCC	0.733	0.670 (±0.149)	0.800 (±0.100)	0.789 (±0.120)	0.658 (±0.249)
-	R-WCC	0.743	0.747 (±0.064)	0.852 (±0.043)	0.850 (±0.045)	0.772 (±0.072)
-	GR	**0.803**	**0.752 (±0.139)**	**0.856 (±0.092)**	**0.849 (±0.115)**	**0.747 (±0.250)**
segment	WCC	0.829	0.328 (±0.024)	0.519 (±0.015)	0.491 (±0.028)	0.418 (±0.031)
-	RCC	0.767	0.294 (±0.073)	0.510 (±0.062)	0.448 (±0.091)	0.363 (±0.140)
-	R-WCC	0.787	0.301 (±0.064)	0.519 (±0.051)	0.457 (±0.076)	0.388 (±0.086)
-	GR	**0.849**	**0.332 (±0.078)**	**0.551 (±0.063)**	**0.493 (±0.091)**	**0.433 (±0.101)**

Tab.: Evaluation of the different strategies on the artificial and UCI datasets.

Experiments

- The results indicate that the **genetic approach (GR)** always gives the best results on the different tested datasets.
- The **worst conflict choice strategy (WCC)** gives good results when the datasets are simple (i.e. `2d-4c-no0` and `2d-4c-no1` with four clusters with two attributes)
- When the datasets are more complex (with ten attributes and from UCI) the **roulette-wheel strategy (R-WCC)** gives better results than worst conflict choice strategy (WCC).
- The last strategy, **random selection (SCC)**, always gives the poorest results, except for one dataset (`10d-4c-no1`).
- These results are consistent as the **genetic resolution strategy (GR)** better explores the search space, and evaluates solutions which are not explored by the other iterative methods.

Conclusion

Summary :

- **Collaborative clustering** is a method which makes **collaborate** different clustering methods
- A **refinement** step is used to modify the results
- Different **strategies** of conflict order resolution have been presented an compared
- We investigated the use of a **genetic algorithm** to optimize the collaboration
- In future work, we plan to investigate **new strategies** of resolution

Conclusion

- Thank you for your attention.
- Time for questions.

contact : `forestier@unistra.fr`
site web : `http://lsiit-cnrs.unistra.fr/fdbt`
google : "*germain forestier*"