Adaptive Clustering Ensemble (ACE)

Geonhee Lee *et al*, *BMC Bioinformatics*, *2019*. | IF=2.511

MODNBIO

International Journal of Machine Learning and Cybernetics (2019) 10:1227–1246
https://doi.org/10.1007/s13042-017-0756-7

ORIGINAL ARTICLE

Clustering ensemble method

Tahani Alqurashi[1] · Wenjia Wang[1]

Received: 28 September 2015 / Accepted: 20 October 2017 / Published online: 16 January 2018

Outline

- Definition of similarity measures
  - cluster similarity
  - membership similarity
  - certain object
  - uncertain object
  - totally certain object
  - totally uncertain object
  - cluster certainty
- The ACE algorithm
  - transformation
  - generating new consensus clusters
  - enforce hard clustering

Definition of similarity measures

1. S_c: The cluster similarity measure between two clusters.
2. S_x: The membership similarity measure.
3. θ_1: The membership matrix, where the columns of this matrix correspond to clusters and the rows correspond to objects.
4. δ: A binary membership value of an object to a particular cluster, $\delta \in \{0, 1\}$.
5. α_1: A threshold for merging clusters, its value is determined based on S_c.
6. α_2: A certainty threshold for placing an object into a cluster, its value is determined based on S_x: Number of clusters in θ_1.
7. C: The set of all the newly formed clusters after the merging process has concluded.
8. P_c: Cluster certainty, only calculated for a newly formed cluster.

Definition of similarity measures

- cluster similarity

We employ the 'set correlation' as a cluster similarity measurement, which measures the overlap between two clusters and takes their size into account. It has been developed in the Relevance-Set Correlation (RSC) [15] model, as this measure is an equivalent of the Pearson correlation in clustering analysis. After some simplification and derivation, it can be represented as follows:

$$S_c(c^q_{j_q}, c^\ell_{j_\ell}) = \frac{|c^q_{j_q} \cap c^\ell_{j_\ell}| - \frac{|c^q_{j_q}||c^\ell_{j_\ell}|}{n}}{\sqrt{|c^q_{j_q}||c^\ell_{j_\ell}|\left(1 - \frac{|c^q_{j_q}|}{n}\right)\left(1 - \frac{|c^\ell_{j_\ell}|}{n}\right)}}$$

$$= \frac{n.CM(c^q_{j_q}, c^\ell_{j_\ell}) - \sqrt{|c^q_{j_q}||c^\ell_{j_\ell}|}}{\sqrt{(n - |c^q_{j_q}|)(n - |c^\ell_{j_\ell}|)}},$$

Definition of similarity measures

- membership similarity

$$\overleftarrow{c}_g = \{c_i + c_j + \cdots + c_r\}$$

$$\overleftarrow{C} = \{\overleftarrow{c}_1, \ldots, \overleftarrow{c}_g, \ldots\}$$

$$S_x(x_i, \overleftarrow{c}_g) = \frac{1}{\max\{\theta_1(x_i, \overleftarrow{C})\}} \theta_1(x_i, \overleftarrow{c}_g),$$

$$\theta_1(x_i, \overleftarrow{c}_g) = \sum_{u=1}^{r} \delta(x_i, c_u)$$

Definition of similarity measures

- certain object

Certain object: An object, x_i, is defined as a certain object if its maximum membership similarity S_x is greater than a pre-set value α_2, i.e.

$$\max(S_x(x_i, \overleftarrow{C})) > \alpha_2.$$

- uncertain object

Uncertain object: An object is defined to be an uncertain object if its maximum membership similarity S_x is less than or equal to α_2, i.e.

$$\max(S_x(x_i, \overleftarrow{C})) <= \alpha_2. \quad (6$$

Definition of similarity measures

- totally certain object

Totally certain object: An object is defined as a totally certain object if its maximum membership similarity S_x is 1.

- totally uncertain object

Totally uncertain object: An object is defined as a totally uncertain object if its maximum membership similarity S_x is 0.

Definition of similarity measures

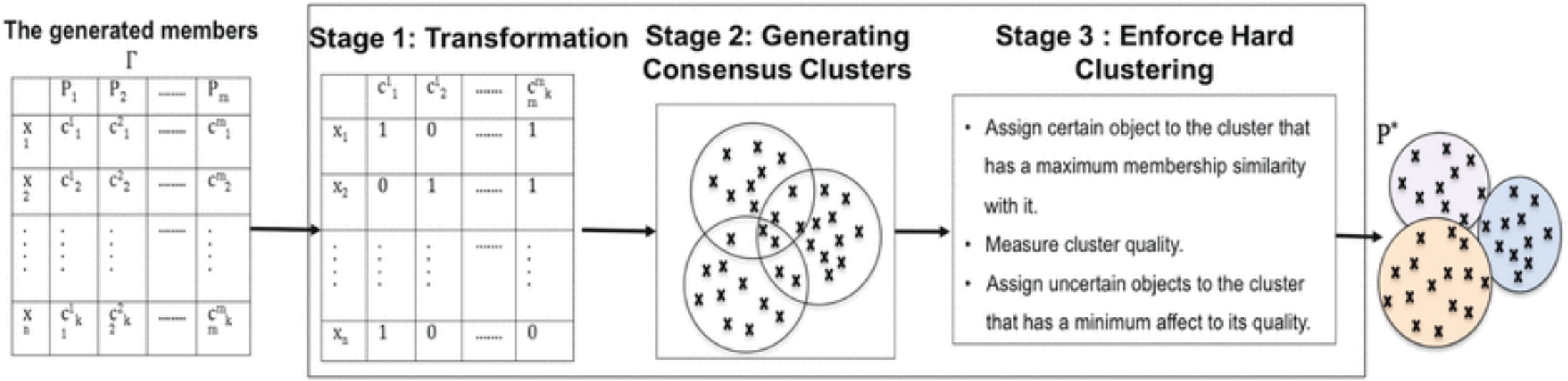
- cluster certainty

Cluster certainty: The cluster certainty, ρ_{c_g}, is defined as the mean of the membership similarity of objects in a cluster $\overleftarrow{c}_g$, i.e.

$$\rho_{c_g} = \frac{1}{|\overleftarrow{c}_g|} \sum_{i=1}^{|\overleftarrow{c}_g|} S_x(x_i, \overleftarrow{c}_g).$$

The cluster certainty is calculated for each newly formed cluster $\in \overleftarrow{C}$.

The ACE algorithm



Definition of similarity measures

- transformation

In general, for cluster c_j in clustering member q, its corresponding vector is represented as $c_j^q = [\delta(x_1), \ldots, \delta(x_n)]^T$, where $\delta(x_i)$ is the binary membership and takes the following value:

$$\delta(x_i, c_j) = \begin{cases} 1, & \text{if } x_i \in c_j, \forall\, i = 1, \ldots, n. \\ 0, & \text{if } x_i \notin c_j \end{cases}$$

Where i is the index of data objects; $j(= 1, \ldots, k_q)$, the index of clusters in each of m members; $q(= 1, \ldots, m)$ is the index of members in an ensemble. There will be k_m vectors to form an $n \times k_m$ cluster matrix $\theta_1 = [c_1^1, c_2^1, \ldots, \ldots, c_{k_m}^q]$. Where $k_m = \sum_{q=1}^{m} k_q$, which is the total number of clusters in all members.

Definition of similarity measures

- generating new consensus clusters

1. *Measuring similarity between initial clusters and merging the most similar ones*

   (a) Starting with k_m initial clusters, we measure the cluster similarity S_c, defined in Eq. 1, between the initial clusters that are placed in different members in Φ.

   (b) The merging process is performed based on the following criterion:

if $S_c(c_{j_q}^q, c_{j_\ell}^\ell) >= \alpha_1 \Rightarrow c_{j_q}^q$ and $c_{j_\ell}^\ell$ are similar, hence merged.

if $S_c(c_{j_q}^q, c_{j_\ell}^\ell) < \alpha_1 \Rightarrow c_{j_q}^q$ and $c_{j_\ell}^\ell$ are dissimilar, not merged.

Definition of similarity measures

- generating new consensus clusters

2. *Producing k consensus clusters*

After the most similar initial clusters are merged, we have θ_1 to represent newly formed clusters and perhaps some remaining non-merged initial clusters. The next step is to check if the number of the clusters in θ_1 is exactly equal to k clusters, which will be taken as the final candidate clusters. For convenience, let λ be the number of clusters in θ_1. There are three possible scenarios: (1) $\lambda = k$, (2) $\lambda > k$, and (3) $\lambda < k$, when checking the number of clusters in θ_1.

(a) When $\lambda = k$, i.e. the number of clusters in θ_1 is equal to the pre-defined k, we then take the clusters in θ_1 as the candidate clusters and adapt α_2 to a value based on S_x so that it can represent a specific percentage of the membership certainty. Then we move onto Stage 3.

Definition of similarity measures

- generating new consensus clusters

(b) When $\lambda > k$, i.e. the number of clusters in θ_1 is greater than the pre-defined k, which is the most likely scenario in practice, there are two options: (A) to terminate the process or (B) to forge ahead with brutal merging or eliminating.

Option A: Coming to this point, the clusters in θ_1 are more dissimilar from each than the given threshold α_1. If the value of α_1 has reached the minimum acceptable similarity, it indicates that the clusters in θ_1 for the given dataset are too dissimilar from each other to be merged to obtain the intended k number of clusters. We then conclude that the pre-set value for k is unreasonable and unachievable, and output the generated clusters.

Option B: However, as there is no gold-standard for setting up the minimum acceptable similarity threshold, it is then also reasonable to go ahead with the process by adapting the threshold value α_1 to reflect the similarity distribution in the current similarity matrix S_c, and then merging the clusters with the above described step, or eliminating the clusters with the following steps. The elimination is carried out based on the cluster certainty. The certainty of each cluster in θ_1 is calculated by Eq. 7 and their certainty values are ranked in a descending order.

Definition of similarity measures

- enforce hard clustering

2. **Identify totally certain and certain objects in θ_1 as in definitions 5 and 3.**

As certain objects have a higher similarity value than α_2, we assign them to the cluster that has a maximum membership similarity among other clusters in θ_1.

Definition of similarity measures

- enforce hard clustering

3. **Measure the quality of each candidate cluster in** θ_1. In principle, any cluster quality measure can be used, so in this study we measure the compactness of the certain objects in a cluster as the quality metric, and here we call it the original quality of each cluster.

The compactness of a cluster is usually measured by the variance, *Var*, which is the average of the squared differences from the mean, as follows:

$$Var(c) = \frac{1}{|\overleftarrow{c}|}\sum_{i=1}^{|\overleftarrow{c}|}(S_x(x_i, \overleftarrow{c}) - p_{\overleftarrow{c}})^2 \quad (11$$

한 클러스터에서 통일성이 고른 정도

Definition of similarity measures

- enforce hard clustering

4. **Identify uncertain objects in θ_1 as in equation 6.**

For each uncertain object the following steps are performed:

(a) Identify the clusters of the current uncertain object in θ_1

(b) For each identified cluster, we recalculate its quality using the Eq. 11 by including the current object membership similarity with the identified cluster.

(c) Compare the original quality and the current quality of the identified clusters.

(d) Assign the current object to the cluster that has a minimum effect on its original quality.

(e) Increase the size of the assigned cluster by 1.

(f) Update the original quality of the assigned cluster to be equal to the current quality.

(g) Repeat steps until all the uncertain objects are assigned.

An illustrating example for the ACE

- Assumption

We illustrate how the ACE works with a simple example. Suppose we have a dataset X that contains 10 objects, $X = \{x_1, x_2, \ldots, x_{10}\}$ and that we have generated 3 members ($m = 3$), each of which has 3 clusters ($k = 3$). We set $\alpha_1 = 0.8$, $\alpha_2 = 0.5$, and $k = 3$, and we run the ACE algorithm in three stages as follows:

An illustrating example for the ACE

- transformation

The generated members

objects	m_1	m_2	m_3
x_1	2	3	1
x_2	2	3	1
x_3	3	1	1
x_4	3	1	3
x_5	3	1	3
x_6	1	3	2
x_7	1	2	2
x_8	1	2	2
x_9	1	3	2
x_{10}	2	2	1

Binary vectors representation of the initial clusters

c_1^1	c_2^1	c_3^1	c_1^2	c_2^2	c_3^2	c_1^3	c_2^3	c_3^3
0	1	0	0	0	1	1	0	0
0	1	0	0	0	1	1	0	0
0	0	1	1	0	0	1	0	0
0	0	1	1	0	0	0	0	1
0	0	1	1	0	0	0	0	1
1	0	0	0	0	1	0	1	0
1	0	0	0	1	0	0	1	0
1	0	0	0	1	0	0	1	0
1	0	0	0	0	1	0	1	0
0	1	0	0	1	0	1	0	0

Fig. 7 An illustrative example of three clustering members for dataset X of 10 objects, and the transformation from members into a binary matrix representation

An illustrating example for the ACE

- generating consensus cluster stage

Table 6 The Similarity Matrix S_c, which is the result of measuring the similarity between initials cluster vectors in our illustrative example (Figure 7) using S_c measure. – – – cells indicates that this similarity is not calculated as they are placed in the same member

	c_1^1	c_2^1	c_3^1	c_1^2	c_2^2	c_3^2	c_1^3	c_2^3	c_3^3
c_1^1	–	–	–	−0.535	0.802	−0.250	−0.667	1	−0.408
c_2^1	–	–	–	−0.429	−0.429	0.802	0.802	−0.535	−0.327
c_3^1	–	–	–	1	−0.429	−0.535	−0.089	−0.535	0.764
c_1^2	−0.535	−0.429	1	–	–	–	−0.089	−0.535	0.764
c_2^2	0.802	−0.429	−0.429	–	–	–	−0.535	0.802	−0.327
c_3^2	−0.250	0.802	−0.535	–	–	–	0.583	−0.250	−0.408
c_1^3	−0.667	0.802	−0.089	−0.089	−0.535	0.583	–	–	–
c_2^3	1	−0.535	−0.535	−0.535	0.802	−0.250	–	–	–
c_3^3	−0.408	−0.327	0.764	0.764	−0.327	−0.408	–	–	–

Table 7 The result of θ_1 after we merge the most similar clusters, which are $\overline{c}_1 = \{c_1^1 + c_2^2 + c_2^3\}$, $\overline{c}_2 = \{c_2^1 + c_3^2 + c_1^3\}$, $\overline{c}_3 = \{c_3^1 + c_1^2\}$ and $\overline{c}_4 = \{c_3^3\}$

	$\overline{c}_1$	$\overline{c}_2$	$\overline{c}_3$	$\overline{c}_4$
x_1	0	3	0	0
x_2	0	3	0	0
x_3	0	1	2	0
x_4	0	0	2	1
x_5	0	0	2	1
x_6	2	1	0	0
x_7	3	0	0	0
x_8	3	0	0	0
x_9	3	0	0	0
x_{10}	0	3	0	0

Table 8 The updated Similarity Matrix S_c after the first step of the merging process is performed, which is the result of measuring the similarity between four clusters in θ_1 (in Table 7)

	$\overline{c}_1$	$\overline{c}_2$	$\overline{c}_3$	$\overline{c}_4$
$\overline{c}_1$	–	−0.408	−0.535	−0.408
$\overline{c}_2$	−0.408	–	−0.218	−0.500
$\overline{c}_3$	−0.535	−0.218	–	0.764
$\overline{c}_4$	−0.408	−0.500	0.764	–

An illustrating example for the ACE

- enforce hard clustering
  - adopting a_1

Table 9 The result of updating θ_1 after we merge $\bar{c}_3$ and $\bar{c}_4$ by summing their objects membership similarity and result in $\bar{c}_3$

	$\bar{c}_1$	$\bar{c}_2$	$\bar{c}_3$
x_1	0	3	0
x_2	0	3	0
x_3	0	1	2
x_4	0	0	3
x_5	0	0	3
x_6	2	1	0
x_7	3	0	0
x_8	3	0	0
x_9	3	0	0
x_{10}	0	3	0

Table 10 The results of S_x after no more merging step is needed

	$\bar{c}_1$	$\bar{c}_2$	$\bar{c}_3$
x_1	0	1	0
x_2	0	1	0
x_3	0	0.33	0.67
x_4	0	0	1
x_5	0	0	1
x_6	0.67	0.33	0
x_7	1	0	0
x_8	1	0	0
x_9	1	0	0
x_{10}	0	1	0

An illustrating example for the ACE

- enforce hard clustering
  - elimination
  - high a_2

Table 12 The result of S_x after we perform the second stage

	$\bar{c}_1$	$\bar{c}_2$	$\bar{c}_3$
x_1	0	1	0
x_2	0	1	0
x_3	0	0.33	0.67
x_4	0	0	0.67
x_5	0	0	0.67
x_6	0.67	0.33	0
x_7	1	0	0
x_8	1	0	0
x_9	1	0	0
x_{10}	0	1	0

An illustrating example for the ACE

- enforce hard clustering
  - elimination
  - high a_2

$$Var(\overleftarrow{c_1}) = \frac{1}{3}((1-0.9)^2 + (1-0.9)^2 + (1-0.9)^2) = 0.01$$
$$Var(\overleftarrow{c_2}) = \frac{1}{3}((1-0.72)^2 + (1-0.72)^2 + (1-0.72)^2) = 0.0784$$
$$Var(\overleftarrow{c_3}) = 0$$

(a) For each candidate cluster we recalculate its quality by including this time x_3:

$$Var(\overleftarrow{c_1}) = \frac{1}{4}((1-0.9)^2 + (1-0.9)^2 + (1-0.9)^2 + (0-0.9)^2) = 0.21$$
$$Var(\overleftarrow{c_2}) = \frac{1}{4}((1-0.72)^2 + (1-0.72)^2 + (1-0.72)^2 + (0.3-0.72)^2) = 0.1029$$
$$Var(\overleftarrow{c_3}) = \frac{1}{1}((0.6-0.6)^2) = 0$$

(b) We compare for each cluster the original quality and the current quality:

$$Var(\overleftarrow{c_1}) = 0.21 - 0.01 = 0.2,$$
$$Var(\overleftarrow{c_2}) = 0.1029 - 0.0784 = 0.0245,$$
$$Var(\overleftarrow{c_3}) = 0 - 0 = 0$$

An illustrating example for the ACE

- enforce hard clustering
  - elimination
  - high a_2

(c) We assign x_3 to the cluster that has a minimum effect on its quality, that is done as follows: $min\{0.2, 0.0245, 0\} = 0$. So, we assign x_3 to cluster $\overleftarrow{c_3}$.
(d) We increase the size of $\overleftarrow{c_3}$ by 1.
(e) We update the original quality of $\overleftarrow{c_3}$ to be equal to the current quality.

After all the uncertain objects are assigned, we produce the final clustering result, which is : $P^* = \{2, 2, 3, 3, 3, 1, 1, 1, 1, 2\}$.