

SDM 2009

9th SIAM International Conference
on Data Mining

Sparks, Nevada, April 30-May 2, 2009

Diversity-based Weighting Schemes for Clustering Ensembles

F. Gullo, A. Tagarelli, S. Greco

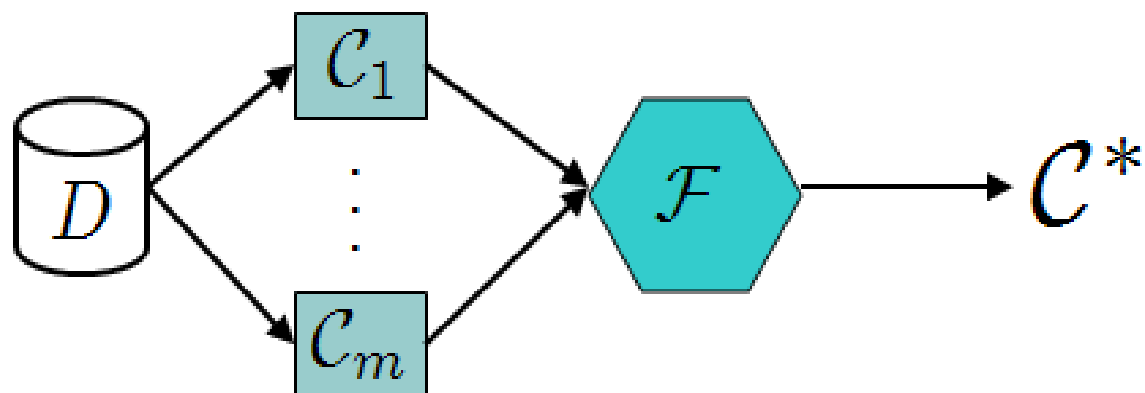
DEIS - University of Calabria - Italy

UNIVERSITÀ DELLA CALABRIA



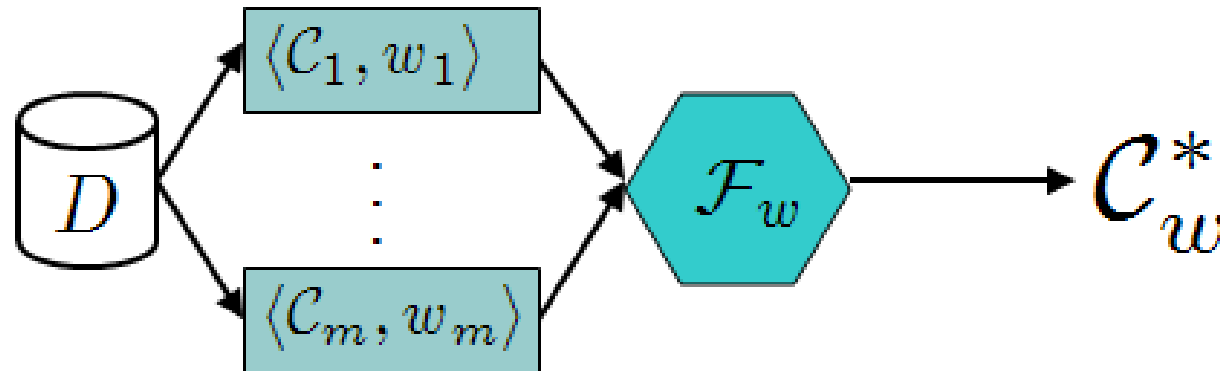
Dipartimento di ELETTRONICA,
INFORMATICA E SISTEMISTICA

Clustering Ensembles (CE)



- each clustering solution C_i is equally considered by the consensus function \mathcal{F}
- three main categories of approaches:
 - ❑ *instance-based CE*
 - ❑ *cluster-based CE*
 - ❑ *hybrid CE*

Weighted Clustering Ensembles (WCE)



- each clustering solution C_i is considered by the consensus function \mathcal{F}_w in a way proportional to the corresponding weight w_i

MAIN ISSUE:

define weights so that **(i)** they are based on some factor that is strongly related to clustering ensembles accuracy, and **(ii)** they are general and easily applicable to any instance-based, cluster-based and hybrid CE method

Diversity-based Weighting Schemes for Clustering Ensembles



Defining *weights*:

- exploits different implementations of the notion of **ensemble diversity**
- takes into account correlations among the individual clustering solutions to different levels



Three proposed weighting schemes:

- *Single Weighting (SW)*
- *Group Weighting (GW)*
- *Dendrogram Weighting (DW)*

Three algorithm schemes designed to easily involve SW, GW and DW into any instance based, cluster-based and hybrid CE algorithm:

- *Weighted instance-based CE (WICE)*
- *Weighted cluster-based CE (WCCE)*
- *Weighted hybrid CE (WHCE)*



Outline

- ❑ Introduction: Clustering Ensembles (*CE*) and Weighted Clustering Ensembles (*WCE*)
- ❑ Background
- ❑ The proposed weighting schemes: *Single Weighting (SW)*, *Group Weighting (GW)*, *Dendrogram Weighting (DW)*
- ❑ Involving weights in CE algorithms: *Weighted instance-based CE (WICE)*, *Weighted cluster-based CE (WCCE)*, *Weighted hybrid CE (WHCE)*
- ❑ Experimental results
- ❑ Conclusions

Background

CLUSTERING SOLUTION. Given a set of data objects D , a *clustering solution* (or *partition*) $\mathcal{C} = \{C_1, \dots, C_k\}$ defined over D is a partition of D into k disjoint groups (clusters)

CLUSTERING ENSEMBLE. Given a set of data objects D , an *ensemble* is a set $E = \{\mathcal{C}_1, \dots, \mathcal{C}_m\}$, where \mathcal{C}_i is a clustering solution defined over D , for each $i \in [1..m]$

CONSENSUS PARTITION. Given a clustering ensemble E , a *consensus partition* derived from E is a clustering solution \mathcal{C}_E^* that maximizes a given *consensus function* by exploiting information available from E



Background

PARTITION-THROUGH DIVERSITY. Given a clustering ensemble E , a *partition-through diversity* measure defined over E is a function $\delta_P : E \times E \rightarrow \mathfrak{R}$ that quantifies, for each pair of clustering solutions $\mathcal{C}_i, \mathcal{C}_j \in E$, how \mathcal{C}_i and \mathcal{C}_j are dissimilar to each other

ENSEMBLE DIVERSITY. Given a clustering ensemble $E = \{\mathcal{C}_1, \dots, \mathcal{C}_m\}$ and a partition-through diversity measure δ_P defined over E , the *ensemble diversity* of E is defined as:

$$\delta_E = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^m \delta_P(\mathcal{C}_i, \mathcal{C}_j)$$

Weighting Clustering Ensembles: *Single Weighting (SW)*

Each clustering solution $\mathcal{C}_i \in E$ is considered individually:

$$W = (w_1, \dots, w_m) = \alpha W' + (1 - \alpha) W'',$$
$$w_i \in [0, 1], \text{ for each } i \in [1..m], \sum_{i=1}^m w_i = 1$$



$W' = (w'_1, \dots, w'_m)$, is a linearly increasing distribution
(*maximum diversity criterion*)

$W'' = (w''_1, \dots, w''_m)$, is a Normal distribution
(*median diversity criterion*)

Weighting Clustering Ensembles: *Single Weighting (SW)*

The components $w'_i \in W'$ increase as $\delta_{E \setminus C_i}$ decreases:

$$w'_i = \left(1 - \frac{\delta_{E \setminus \{C_i\}}}{\sum_{l=1}^m \delta_{E \setminus \{C_l\}}} \right) / (m - 1)$$

The components $w''_i \in W''$ are defined so that the maximum value in W'' corresponds to the clustering solution C_i having the median $\delta_{E \setminus C_i}$:

$$w''_i = \frac{N_{\mu, \sigma}(\delta_{E \setminus \{C_i\}})}{\sum_{l=1}^m N_{\mu, \sigma}(\delta_{E \setminus \{C_l\}})}$$

where $N_{\mu, \sigma}$ is the Normal probability density function having mean μ and standard deviation σ



Weighting Clustering Ensembles: *Single Weighting (SW)*

MAIN ISSUE:

the clustering solutions in the ensemble are considered individually; however, an ensemble may not contain only solutions that are totally dissimilar to each other...

Weighting Clustering Ensembles: *Group Weighting (GW)*

1. partition the ensemble E into a set of clusters (of clusterings)
 $\mathbf{C} = \{C_1, \dots, C_k\}$
2. compute the vector $W_{\mathbf{C}} = (w_{\mathbf{C}}^{(1)}, \dots, w_{\mathbf{C}}^{(k)})$ of *macro*-weights, where each $w_{\mathbf{C}}^{(l)}$, $l \in [1..k]$, is assigned to the cluster (of clusterings) $C_l \in \mathbf{C}$
3. compute the vector $W = (w_1, \dots, w_m)$ of *micro*-weights from $W_{\mathbf{C}}$, in which each w_i , $i \in [1..m]$, is assigned to the clustering solution $C_i \in E$

Weighting Clustering Ensembles: *Group Weighting (GW)*

Computing *macro*-weights:

$$\begin{aligned} W_{\mathbf{C}} &= \alpha W'_{\mathbf{C}} + (1 - \alpha) W''_{\mathbf{C}} = \\ &= \alpha \left(w_{\mathbf{C}}^{(1)'} , \dots , w_{\mathbf{C}}^{(k)'} \right) + (1 - \alpha) \left(w_{\mathbf{C}}^{(1)''} , \dots , w_{\mathbf{C}}^{(k)''} \right) \end{aligned}$$



$$w_{\mathbf{C}}^{(l)'} = \left(1 - \frac{\delta_{E \setminus C_l}}{\sum_{u=1}^k \delta_{E \setminus C_u}} \right) / (k - 1)$$

$$w_{\mathbf{C}}^{(l)''} = \frac{N_{\mu, \sigma}(\delta_{E \setminus C_l})}{\sum_{u=1}^k N_{\mu, \sigma}(\delta_{E \setminus C_u})}$$

Weighting Clustering Ensembles: *Group Weighting (GW)*

Computing *micro*-weights:

$$w_i = w_{i,j}^{SW} \times \bar{w}_j$$



$w_{i,j}^{SW}$ is the weight assigned to the clustering solution \mathcal{C}_i according to the SW scheme, when the ensemble is given by $\mathcal{C}_j \in \mathbf{C}$. \mathcal{C}_j is the cluster such that $\mathcal{C}_i \in \mathcal{C}_j$

$\bar{w}_j \in W_{\mathbf{C}}$ is the weight assigned to \mathcal{C}_j in the first step of GW



Weighting Clustering Ensembles: *Group Weighting (GW)*

MAIN ISSUE:

GW requires a clustering algorithm to partition the ensemble and its relative parameter settings, such as the number of output clusters...

Weighting Clustering Ensembles: *Dendrogram Weighting (DW)*

LEVEL-ORGANIZED DENDROGRAM.

A *level-organized dendrogram* is a set $\mathcal{D} = \{\mathcal{L}_0, \dots, \mathcal{L}_\tau\}$, where each \mathcal{L}_u , $u \in [0..\tau]$, is a set of clusters $\{C_1^{(u)}, \dots, C_{k_u}^{(u)}\}$ corresponding to the level u , such that:

1. $\bigcup_{v=1}^{k_u} C_v^{(u)} = D$
2. $C_v^{(u)} \cap C_w^{(u)} = \emptyset, \forall C_v^{(u)}, C_w^{(u)} \in \mathcal{L}_u$
3. $|\mathcal{L}_0| = |D|, |\mathcal{L}_\tau| = 1, |\mathcal{L}_u| > |\mathcal{L}_{u+1}|, u \in [0..\tau-1]$

Weighting Clustering Ensembles: *Dendrogram Weighting (DW)*

The weight vector W is computed by associating each clustering solution $\mathcal{C}_i \in E$ with a coefficient γ_i :

$$\gamma_i = \sum_{h=1}^{\tau-1} (\tau - h) I(\mathcal{D}, \mathcal{C}_i, \mathcal{L}_h)$$

$$I(\mathcal{D}, \mathcal{C}_i, \mathcal{L}_h) = \begin{cases} 1 & \text{if } \bar{C}_h \neq \bar{C}_{h-1} \\ 0 & \text{otherwise} \end{cases} \quad \text{where } \bar{C}_h \in \mathcal{L}_h \text{ and } \bar{C}_{h-1} \in \mathcal{L}_{h-1}$$

are the clusters such that $\mathcal{C}_i \in \bar{C}_h$ and $\mathcal{C}_i \in \bar{C}_{h-1}$

W is finally computed by applying the same equations used for SW, where $\delta_{E \setminus \{\mathcal{C}_i\}}$ is replaced with the coefficients γ_i



Weighting Clustering Ensembles: *Dendrogram Weighting (DW)*

γ_i expresses the correlation of \mathcal{C}_i with the other clusterings in the ensemble, based on the set S_i , i.e., the set of different clusters of the dendrogram that contain \mathcal{C}_i :

1. the higher the correlation of \mathcal{C}_i with the other clusterings, the higher γ_i
2. γ_i is directly proportional to the size of S_i
3. γ_i is inversely proportional to the sum of dendrogram levels that contain the clusters in S_i

Involving weights in CE algorithms: *Weighted instance-based CE (WICE)*

Algorithm 1 WICE: Weighted Instance-based Clustering Ensembles

Input: a set of data objects $D = \{x_1, \dots, x_n\}$, where
 $x_j = (f_{j1}, \dots, f_{jp})$, $j \in [1..n]$;
an ensemble $E = \{C_1, \dots, C_m\}$ defined over D ;
a weight vector $W = (w_1, \dots, w_m)$

Output: the consensus partition C_E^*

- 1: **for all** $j \in [1..n]$ **do**
- 2: replace x_j with $x'_j = (f'_{j1}, \dots, f'_{jm})$
- 3: **end for**
- 4: **for all** $a \in [1..n]$, $b \in [1..n]$ **do**
- 5: $(M')_{ab} = \Phi(w_1\phi(f'_{a1}, f'_{b1}), \dots, w_m\phi(f'_{am}, f'_{bm}), \Gamma(x'_a, x'_b))$
- 6: **end for**
- 7: $C_E^* \leftarrow cluster(D, M')$

Involving weights in CE algorithms: *Weighted cluster-based CE (WCCE)*

Algorithm 2 WCCE: Weighted Cluster-based Clustering Ensembles

Input: a set of data objects $D = \{x_1, \dots, x_n\}$;
an ensemble $E = \{C_1, \dots, C_m\}$ defined over D ;
a weight vector $W = (w_1, \dots, w_m)$

Output: the consensus partition C_E^*

- 1: $C_E^* = \{C_1^*, \dots, C_k^*\} \leftarrow \{\emptyset, \dots, \emptyset\}$
- 2: $D_M \leftarrow \bigcup_{C \in E} C$
- 3: $M_{D_M} \leftarrow \text{pair-wise-distances}(D_M)$
- 4: $M = \{M_1, \dots, M_k\} \leftarrow \text{cluster}(D_M, M_{D_M})$
- 5: **for all** $j \in [1..n]$ **do**
- 6: find $M_l \in M$ such that $M_l \leftarrow \text{assign}(x_j, M, W)$
- 7: $C_l^* \leftarrow C_l^* \cup \{x_j\}$
- 8: **end for**

Involving weights in CE algorithms: *Weighted hybrid CE (WHCE)*

Algorithm 3 WHCE: Weighted Hybrid Clustering Ensembles

Input: a set of data objects $D = \{x_1, \dots, x_n\}$;
an ensemble $E = \{C_1, \dots, C_m\}$ defined over D ;
a weight vector $W = (w_1, \dots, w_m)$

Output: the consensus partition C_E^*

- 1: $\mathcal{V}_o \leftarrow D$
- 2: $\mathcal{V}_c \leftarrow \bigcup_{C \in E} C$
- 3: $\mathcal{E} \leftarrow \emptyset$
- 4: **for all** $v_o \in \mathcal{V}_o$ **do**
- 5: **for all** $v_c \in \mathcal{V}_c$ **do**
- 6: $\omega = \text{weight}(v_o, v_c, E, W)$
- 7: $\mathcal{E} \leftarrow \mathcal{E} \cup \{(v_o, v_c, \omega)\}$
- 8: **end for**
- 9: **end for**
- 10: $G_H \leftarrow \langle \mathcal{V}_o \cup \mathcal{V}_c, \mathcal{E} \rangle$
- 11: $C_E^* \leftarrow \text{partition}(G_H)$



Experiments

The experiments were designed to evaluate the accuracy of the various CE algorithms employing the proposed weighting schemes SW, GW and DW w.r.t. the case where no weight was used

CE algorithms considered in the evaluation:

- ❑ Instance-based:
CSPA, HGPA, WSPA, MV, AGGL, IVC
- ❑ Cluster-based:
MCLA, MCS
- ❑ Hybrid:
HBGF, WBPA

Experiments: *datasets*

<i>dataset</i>	<i>objects</i>	<i>attributes</i>	<i>classes</i>
Glass	214	10	7
Ecoli	327	7	5
ImageSegmentation	2,310	19	7
ISOLET	1,800	617	6
LetterRecognition	7,000	16	10
Tracedata	200	275	4
ControlChart	600	60	6

Glass, Ecoli, ImageSegmentation, ISOLET and LetterRecognition are from UCI Machine Learning Repository

Tracedata and ControlChart are benchmark time series datasets



Experiments: *results*

- ❑ Evaluation of weighted clustering ensembles
 - ❑ *max-avg-min improvements equal to 24%-16%-8%*
- ❑ Evaluation of weighting schemes
 - ❑ *DW led to the maximum accuracy improvements*
 - ❑ *GW led to better maximum performance than SW*
 - ❑ *SW behaved more reliably than GW*

Experiments: *results*

Results on LetterRecognition

<i>method</i>	<i>diversity</i>	<i>accuracy</i>			
		<i>no weights</i>	<i>SW</i>	<i>GW</i>	<i>DW</i>
CSPA	NMI	0.40	0.48	0.47	0.48
	FM	0.51	0.60	0.61	0.62
HPGA	NMI	0.41	0.40	0.38	0.41
	FM	0.51	0.48	0.52	0.53
WSPA	NMI	0.43	0.45	0.45	0.45
	FM	0.52	0.53	0.49	0.53
MV	NMI	0.72	0.70	0.68	0.70
	FM	0.80	0.80	0.80	0.80
AGGL	NMI	0.63	0.64	0.63	0.65
	FM	0.68	0.70	0.74	0.74
IVC	NMI	0.38	0.43	0.41	0.43
	FM	0.46	0.56	0.55	0.58
MCLA	NMI	0.45	0.49	0.51	0.53
	FM	0.56	0.59	0.59	0.62
MCS	NMI	0.48	0.50	0.50	0.53
	FM	0.50	0.52	0.48	0.55
HBGF	NMI	0.40	0.41	0.42	0.42
	FM	0.51	0.52	0.50	0.55
WBPA	NMI	0.46	0.48	0.50	0.51
	FM	0.52	0.52	0.56	0.57

Results on Tracedata

<i>method</i>	<i>diversity</i>	<i>accuracy</i>			
		<i>no weights</i>	<i>SW</i>	<i>GW</i>	<i>DW</i>
CSPA	NMI	0.50	0.51	0.48	0.50
	FM	0.53	0.54	0.51	0.54
HPGA	NMI	0.53	0.55	0.56	0.58
	FM	0.64	0.65	0.67	0.67
WSPA	NMI	0.50	0.50	0.50	0.50
	FM	0.52	0.55	0.55	0.57
MV	NMI	0.50	0.54	0.57	0.54
	FM	0.53	0.59	0.62	0.63
AGGL	NMI	0.50	0.57	0.58	0.57
	FM	0.54	0.64	0.60	0.64
IVC	NMI	0.50	0.58	0.56	0.59
	FM	0.54	0.63	0.60	0.64
MCLA	NMI	0.58	0.60	0.63	0.64
	FM	0.71	0.70	0.73	0.75
MCS	NMI	0.57	0.58	0.57	0.60
	FM	0.63	0.68	0.65	0.66
HBGF	NMI	0.50	0.51	0.53	0.54
	FM	0.53	0.60	0.62	0.62
WBPA	NMI	0.45	0.50	0.53	0.56
	FM	0.52	0.53	0.56	0.62