Equitable Conceptual Clustering using OWA operator

N. Aribi[1], A. Ouali[2], Y. Lebbah[1], **S. Loudni**[2]

(1) University of Oran 1, LITIO, Oran, Algeria

(2) University of Caen Normandy, GREYC (CNRS UMR 6072), Caen, France

Journée Normastic, 4 Juillet 2018

Outline

Conceptual clustering

Input

- $\mathcal{I}$ set of n distinct literals (items)
- $\mathcal{T}$ multi-set (dataset) of m transactions t (itemsets) s.t. $t \subseteq \mathcal{I}$
- R binary relationship between $\mathcal{T}$ and $\mathcal{I}$ s.t. $(t, i) \in R$ iff $i \in t$

Trans.	Items							
t_1	A	B		D				
t_2	A				E	F		
t_3	A				E		G	
t_4	A				E		G	
t_5		B			E		G	
t_6		B			E		G	
t_7			C		E		G	
t_8			C		E		G	
t_9			C		E			H
t_{10}			C		E			H
t_{11}			C			F	G	H

Conceptual clustering

Input

- $\mathcal{I}$ set of n distinct literals (items)
- $\mathcal{T}$ multi-set (dataset) of m transactions t (itemsets) s.t. $t \subseteq \mathcal{I}$
- R binary relationship between $\mathcal{T}$ and $\mathcal{I}$ s.t. $(t, i) \in R$ iff $i \in t$

Extent of $I \subseteq \mathcal{I}$, $ext(I) = \{t \in \mathcal{T} | \forall i \in I, (t, i) \in R\}$

Trans.	Items							
t_1	A	B		D				
t_2	A				E	F		
t_3	A				E		G	
t_4	A				E		G	
t_5		B			E		G	
t_6		B			E		G	
t_7			C		E		G	
t_8			C		E		G	
t_9			C		E			H
t_{10}			C		E			H
t_{11}			C			F	G	H

$ext(\{B, E, G\}) = \{t_5, t_6\}$

Conceptual clustering

Input

- $\mathcal{I}$ set of n distinct literals (items)
- $\mathcal{T}$ multi-set (dataset) of m transactions t (itemsets) s.t. $t \subseteq \mathcal{I}$
- R binary relationship between $\mathcal{T}$ and $\mathcal{I}$ s.t. $(t, i) \in R$ iff $i \in t$

Intent of $T \subseteq \mathcal{T}$, $int(T) = \{i \in \mathcal{I} | \; \forall t \in T, (t, i) \in R\}$

Trans.	Items							
t_1	A	B		D				
t_2	A				E	F		
t_3	A				E		G	
t_4	A				E		G	
t_5		B			E		G	
t_6		B			E		G	
t_7			C		E		G	
t_8			C		E		G	
t_9			C		E			H
t_{10}			C		E			H
t_{11}			C			F	G	H

$int(\{t_5, t_6\}) = \{B, E, G\}$

Conceptual clustering

Input

- $\mathcal{I}$ set of n distinct literals (items)
- $\mathcal{T}$ multi-set (dataset) of m transactions t (itemsets) s.t. $t \subseteq \mathcal{I}$
- R binary relationship between $\mathcal{T}$ and $\mathcal{I}$ s.t. $(t, i) \in R$ iff $i \in t$

itemset = a set of items

Trans.	Items							
t_1	A	B		D				
t_2	A				E	F		
t_3	A				E		G	
t_4	A				E		G	
t_5		B			E		G	
t_6		B			E		G	
t_7			C		E		G	
t_8			C		E		G	
t_9			C		E			H
t_{10}			C		E			H
t_{11}			C			F	G	H

Conceptual clustering

Input

- $\mathcal{I}$ set of n distinct literals (items)
- $\mathcal{T}$ multi-set (dataset) of m transactions t (itemsets) s.t. $t \subseteq \mathcal{I}$
- R binary relationship between $\mathcal{T}$ and $\mathcal{I}$ s.t. $(t, i) \in R$ iff $i \in t$

Formal concept : a couple (T, I) s.t. $I = int(T) \wedge T = ext(I)$

Trans.	Items							
t_1	A	B		D				
t_2	A				E	F		
t_3	A				E		G	
t_4	A				E		G	
t_5		B			E		G	
t_6		B			E		G	
t_7			C		E		G	
t_8			C		E		G	
t_9			C		E			H
t_{10}			C		E			H
t_{11}			C			F	G	H

$\phi = (\{t_5, t_6\}, \{B, E, G\})$

- $ext(\{B, E, G\}) = \{t_5, t_6\}$
- $int(\{t_5, t_6\}) = \{B, E, G\}$
- $\mathtt{freq}(\{B, E, G\}) = 2$

Conceptual clustering

Input

- $\mathcal{I}$ set of n distinct literals (items)
- $\mathcal{T}$ multi-set (dataset) of m transactions t (itemsets) s.t. $t \subseteq \mathcal{I}$
- R binary relationship between $\mathcal{T}$ and $\mathcal{I}$ s.t. $(t, i) \in R$ iff $i \in t$

Formal concept : a couple (T, I) s.t. $I = int(T) \wedge T = ext(I)$

Trans.	Items							
t_1	A	B		D				
t_2	A				E	F		
t_3	A				E		G	
t_4	A				E		G	
t_5		B			E		G	
t_6		B			E		G	
t_7			C		E		G	
t_8			C		E		G	
t_9			C		E			H
t_{10}			C		E			H
t_{11}			C			F	G	H

- Clustering : Partition of $\mathcal{T}$
- Conceptual clustering : Each cluster is a formal concept

Conceptual Clustering

Conceptual clustering $\equiv$ set of k **formal concepts** $\Phi = \{\phi_1, \ldots, \phi_k\}$, where $\phi_j = (I_j, T_j)$, such that $\{T_1, \ldots, T_k\}$ forms a **partition** of the set of transactions $\mathcal{T}$.

$$(Q) \begin{cases} k_{min} \leq k \leq k_{max} \ \wedge \\ \bigcup_{i \in [1..k]} T_i = \mathcal{T} \ \wedge \\ \bigwedge_{i,j \in [1..k]} T_i \cap T_j = \emptyset \ \wedge \\ \bigwedge_{j \in [1..k]} \texttt{closed}(\phi_j) \end{cases}$$

Conceptual Clustering

Conceptual clustering $\equiv$ set of k **formal concepts** $\Phi = \{\phi_1, \ldots, \phi_k\}$, where $\phi_j = (I_j, T_j)$, such that $\{T_1, \ldots, T_k\}$ forms a **partition** of the set of transactions $\mathcal{T}$.

$$
(Q) \begin{cases} k_{min} \leq k \leq k_{max} \wedge \\ \bigcup_{i \in [1..k]} T_i = \mathcal{T} \wedge \\ \bigwedge_{i,j \in [1..k]} T_i \cap T_j = \emptyset \wedge \\ \bigwedge_{j \in [1..k]} \texttt{closed}(\phi_j) \end{cases}
$$

- ➠ $k = size(\Phi) = |\Phi|$
- ➠ all transactions are **covered**
- ➠ **without any overlap** between clusters
- ➠ each formal concept is a **closed itemset**

Conceptual Clustering

Conceptual clustering $\equiv$ set of k **formal concepts** $\Phi = \{\phi_1, \ldots, \phi_k\}$, where $\phi_j = (I_j, T_j)$, such that $\{T_1, \ldots, T_k\}$ forms a **partition** of the set of transactions $\mathcal{T}$.

$$(Q) \begin{cases} k_{min} \leq k \leq k_{max} \; \wedge \\ \bigcup_{i \in [1..k]} T_i = \mathcal{T} \; \wedge \\ \bigwedge_{i,j \in [1..k]} T_i \cap T_j = \emptyset \; \wedge \\ \bigwedge_{j \in [1..k]} \mathtt{closed}(\phi_j) \end{cases}$$

- ➟ $k = size(\Phi) = |\Phi|$
- ➟ all transactions are **covered**
- ➟ **without any overlap** between clusters
- ➟ each formal concept is a **closed itemset**

- $k = 3$

Trans.	Items							
t_1	A	B		D				
t_2	A				E	F		
t_3	A				E		G	
t_4	A				E		G	
t_5		B			E		G	
t_6		B			E		G	
t_7			C		E		G	
t_8			C		E		G	
t_9			C		E			H
t_{10}			C		E			H
t_{11}			C			F	G	H

Sol.	X_1	X_2	X_3
s_1	{C, F, G, H}	{E}	{A, B, D}

Conceptual Clustering

Conceptual clustering $\equiv$ set of k **formal concepts** $\Phi = \{\phi_1, \ldots, \phi_k\}$, where $\phi_j = (I_j, T_j)$, such that $\{T_1, \ldots, T_k\}$ forms a **partition** of the set of transactions $\mathcal{T}$.

$$(Q) \begin{cases} k_{min} \leq k \leq k_{max} \ \wedge \\ \bigcup_{i \in [1..k]} T_i = \mathcal{T} \ \wedge \\ \bigwedge_{i,j \in [1..k]} T_i \cap T_j = \emptyset \ \wedge \\ \bigwedge_{j \in [1..k]} \texttt{closed}(\phi_j) \end{cases}$$

- ➠ $k = size(\Phi) = |\Phi|$
- ➠ all transactions are **covered**
- ➠ **without any overlap** between clusters
- ➠ each formal concept is a **closed itemset**

- $k = 3$

Trans.	Items							
t_1	A	B		D				
t_2	A				E	F		
t_3	A				E		G	
t_4	A				E		G	
t_5		B			E		G	
t_6		B			E		G	
t_7			C		E		G	
t_8			C		E		G	
t_9			C		E			H
t_{10}			C		E			H
t_{11}			C			F	G	H

Sol.	X_1	X_2	X_3
s_1	{C, F, G, H}	{E}	{A, B, D}
s_2	{B}	{C}	{A, E}

Conceptual Clustering

Conceptual clustering $\equiv$ set of k **formal concepts** $\Phi = \{\phi_1, \ldots, \phi_k\}$, where $\phi_j = (I_j, T_j)$, such that $\{T_1, \ldots, T_k\}$ forms a **partition** of the set of transactions $\mathcal{T}$.

$$(Q)\begin{cases} k_{min} \leq k \leq k_{max} \;\wedge \\ \bigcup_{i \in [1..k]} T_i = \mathcal{T} \;\wedge \\ \bigwedge_{i,j \in [1..k]} T_i \cap T_j = \emptyset \;\wedge \\ \bigwedge_{j \in [1..k]} \texttt{closed}(\phi_j) \end{cases}$$

- ➠ $k = size(\Phi) = |\Phi|$
- ➠ all transactions are **covered**
- ➠ **without any overlap** between clusters
- ➠ each formal concept is a **closed itemset**

- $k = 3$

Trans.	Items							
t_1	A	B		D				
t_2	A				E	F		
t_3	A				E		G	
t_4	A				E		G	
t_5		B			E		G	
t_6		B			E		G	
t_7			C		E		G	
t_8			C		E		G	
t_9			C		E			H
t_{10}			C		E			H
t_{11}			C			F	G	H

Sol.	X_1	X_2	X_3
s_1	{C, F, G, H}	{E}	{A, B, D}
s_2	{B}	{C}	{A, E}
s_3	{A}	{C}	{B, E, G}

Conceptual Clustering

Conceptual clustering $\equiv$ set of k **formal concepts** $\Phi = \{\phi_1, \ldots, \phi_k\}$, where $\phi_j = (I_j, T_j)$, such that $\{T_1, \ldots, T_k\}$ forms a **partition** of the set of transactions $\mathcal{T}$.

$$(Q) \begin{cases} k_{min} \leq k \leq k_{max} \ \wedge \\ \bigcup_{i \in [1..k]} T_i = \mathcal{T} \ \wedge \\ \bigwedge_{i,j \in [1..k]} T_i \cap T_j = \emptyset \ \wedge \\ \bigwedge_{j \in [1..k]} \texttt{closed}(\phi_j) \end{cases}$$

- ➠ $k = size(\Phi) = |\Phi|$
- ➠ all transactions are **covered**
- ➠ **without any overlap** between clusters
- ➠ each formal concept is a **closed itemset**

- $k = 3$

Trans.	Items							
t_1	A	B		D				
t_2	A				E	F		
t_3	A				E		G	
t_4	A				E		G	
t_5		B			E		G	
t_6		B			E		G	
t_7			C		E		G	
t_8			C		E		G	
t_9			C		E			H
t_{10}			C		E			H
t_{11}			C			F	G	H

Sol.	X_1	X_2	X_3
s_1	{C, F, G, H}	{E}	{A, B, D}
s_2	{B}	{C}	{A, E}
s_3	{A}	{C}	{B, E, G}

- **Necessity for optimizing.**
- Maximizing the sum of size provides s_1 as optimal solution (value 8).

ILP model for conceptual clustering (Ouali et al., IJCAI'16)

Two steps approach :

1. Extracting the set $\mathcal{C}$ of all formal concepts by a dedicated closed itemset mining tool
2. Computing an optimal clustering that is a partition of $\mathcal{T}$ using `ILP`

`Optimize`	$z = \sum_{c=1}^{\lvert\mathcal{C}\rvert} v_c \,.\, x_c$
`Under constraints`	(1) $\sum_{c=1}^{\lvert\mathcal{C}\rvert} a_{t,c} \,.\, x_c = 1, \quad \forall t \in \mathcal{T}$
	(2) $\sum_{c=1}^{\lvert\mathcal{C}\rvert} x_c = k$
	(2′) $k_{min} \leq k \leq k_{max}$
	$k \in \mathbb{N}^*, \quad x_c \in \{0, 1\}, c \in \mathcal{C}$

- $(x_c = 1)$ iff closed itemset c **belongs** to clustering
- v_c: value of **quality measure** for closed itemset c: size, diversity, ...
- $(a_{t,c})$ boolean matrix: $(a_{t,c} = 1)$ iff closed itemset c **covers** transaction t
- **Number of clusters** k is not fixed a priori (2), but user-constrained (2')

Motivations: come back to the example

Trans.	Items							
t_1	A	B		D				
t_2	A				E	F		
t_3	A				E		G	
t_4	A				E		G	
t_5		B			E		G	
t_6		B			E		G	
t_7			C		E		G	
t_8			C		E		G	
t_9			C		E			H
t_{10}			C		E			H
t_{11}			C			F	G	H

Sol.	X_1	X_2	X_3
s_1	{C, F, G, H}	{E}	{A, B, D}
s_2	{B}	{C}	{A, E}
s_3	{A}	{C}	{B, E, G}

- Result interpretation

Motivations: come back to the example

Trans.	Items							
t_1	A	B		D				
t_2	A				E	F		
t_3	A				E		G	
t_4	A				E		G	
t_5		B			E		G	
t_6		B			E		G	
t_7			C		E		G	
t_8			C		E		G	
t_9			C		E			H
t_{10}			C		E			H
t_{11}			C			F	G	H

Sol.	X_1	X_2	X_3
s_1	{C, F, G, H}	{E}	{A, B, D}
s_2	{B}	{C}	{A, E}
s_3	{A}	{C}	{B, E, G}

- Result interpretation
  - ➠ clusterings with high frequency but low size
  - ➠ clusterings with high description size but low frequency, many clusters

Motivations: come back to the example

Trans.	Items							
t_1	A	B		D				
t_2	A				E	F		
t_3	A				E		G	
t_4	A				E		G	
t_5		B			E		G	
t_6		B			E		G	
t_7			C		E		G	
t_8			C		E		G	
t_9			C		E			H
t_{10}			C		E			H
t_{11}			C			F	G	H

Sol.	X_1	X_2	X_3
s_1	{C, F, G, H}	{E}	{A, B, D}
s_2	{B}	{C}	{A, E}
s_3	{A}	{C}	{B, E, G}

- Result interpretation
- **Maximizing the sum of sizes of selected concepts** : one optimal solution $s_1 = (1, 1, 9)$, cost 8
  ➠ unbalanced solution

Motivations: come back to the example

Trans.	Items							
t_1	A	B		D				
t_2	A				E	F		
t_3	A				E		G	
t_4	A				E		G	
t_5		B			E		G	
t_6		B			E		G	
t_7			C		E		G	
t_8			C		E		G	
t_9			C		E			H
t_{10}			C		E			H
t_{11}			C			F	G	H

Sol.	X_1	X_2	X_3
s_1	{C, F, G, H}	{E}	{A, B, D}
s_2	{B}	{C}	{A, E}
s_3	{A}	{C}	{B, E, G}

- Result interpretation
- **Maximizing the sum of sizes of selected concepts** : one optimal solution $s_1 = (1, 1, 9)$, cost 8
- A more *balanced* clustering: $s_2 = (3, 5, 3)$, cost 4

Motivations: come back to the example

Trans.	Items							
t_1	A	B		D				
t_2	A				E	F		
t_3	A				E		G	
t_4	A				E		G	
t_5		B			E		G	
t_6		B			E		G	
t_7			C		E		G	
t_8			C		E		G	
t_9			C		E			H
t_{10}			C		E			H
t_{11}			C			F	G	H

Sol.	X_1	X_2	X_3
s_1	{C, F, G, H}	{E}	{A, B, D}
s_2	{B}	{C}	{A, E}
s_3	{A}	{C}	{B, E, G}

- Result interpretation
- **Maximizing the sum of sizes of selected concepts** : one optimal solution $s_1 = (1, 1, 9)$, cost 8
- A more *balanced* clustering: $s_2 = (3, 5, 3)$, cost 4
- **Dedicated optimization settings** to obtain more balanced clusterings
  - *maximizing the minimal frequency* (**Maxmin**)
  - *minimizing the deviation in cluster frequency* (**minDev**)

Motivations: come back to the example

Trans.	Items							
t_1	A	B		D				
t_2	A				E	F		
t_3	A				E		G	
t_4	A				E		G	
t_5		B			E		G	
t_6		B			E		G	
t_7			C		E		G	
t_8			C		E		G	
t_9			C		E			H
t_{10}			C		E			H
t_{11}			C			F	G	H

Sol.	X_1	X_2	X_3
s_1	{C, F, G, H}	{E}	{A, B, D}
s_2	{B}	{C}	{A, E}
s_3	{A}	{C}	{B, E, G}

- Result interpretation

- **Maximizing the sum of sizes of selected concepts** : one optimal solution $s_1 = (1, 1, 9)$, cost 8

- A more *balanced* clustering: $s_2 = (3, 5, 3)$, cost 4

- **Dedicated optimization settings** to obtain more balanced clusterings
  - *maximizing the minimal frequency* (**Maxmin**)
  - *minimizing the deviation in cluster frequency* (**minDev**)
    ➠ drowning effect problem : $(1, 1, 1, 100)$ and $(100, 100, 100, 1)$ are indistinguishable

Equity in multi-agent optimization

Let P a multi-agent combinatorial problem

- n agents $N = \{1, .., n\}$
- solution of P $\Leftrightarrow$ utility vector $x = (x_1, ..., x_n) \in R^n_+$ (x_i = the cost of solution x w.r.t. to agent i)
- comparison of solutions reduces to the comparison of their utility vectors

Fairness in area of combinatorial optimization problems refers to:

- fair distribution of utility values among agents
- equitably efficient solutions (i.e. specific refinement of the Pareto-optimality)

.

Widely studied in the context of optimization

- fair resource allocation problem (Bouveret et al. AAMAS'05)
- conference paper assignment problem. (Goldsmith et al. AAAI'07, Lian et al. AAAI'18)
- ...

Formalization of the equity principle

Let $\succsim_{\|}$ be a preference relation on utility vectors. To capture both efficiency and equity in comparisons, $\succsim_{\|}$ should satisfy three main properties:

Formalization of the equity principle

Let $\succsim_{\|}$ be a preference relation on utility vectors. To capture both efficiency and equity in comparisons, $\succsim_{\|}$ should satisfy three main properties:

Symmetry formalizes the fact that all agents are treated equivalently:

For any $x \in R^n_+$, permutation σ on N, we have $(x_{\sigma(1)}, ..., x_{\sigma(n)}) \sim (x_1, ..., x_n)$.

➠ both utility vectors (5,3,0) et (0,3,5) are considered equivalent.

Formalization of the equity principle

Let $\succsim_{\|}$ be a preference relation on utility vectors. To capture both efficiency and equity in comparisons, $\succsim_{\|}$ should satisfy three main properties:

- **P-Monotony** enforces consistency with P-dominance:

For all $x, y \in R_+^n$, $x \succsim_P y \Rightarrow x \succsim_{\|} y$ and $x \succ_P y \Rightarrow x \succ_{\|} y$.

➠ $(\mathbf{5}, 5, 1)$ dominates $(\mathbf{3}, 5, 1)$ (Pareto dominance)

Formalization of the equity principle

Let $\succsim_{\|}$ be a preference relation on utility vectors. To capture both efficiency and equity in comparisons, $\succsim_{\|}$ should satisfy three main properties:

- **Transfer principle (a.k.a Pigou-Dalton transfers)** captures the principle of "fairness" :

Let $x = (x_1, x_2, \ldots, x_i, \ldots, x_j, \ldots, x_n) \in R^n_+$ s.t. $x_i > x_j$ for some i and j. Then for all ϵ s.t. $0 < \epsilon \leq \frac{x_i - x_j}{2}$, $x - \epsilon e_i + \epsilon e_j \succ x$ where e_i (resp. e_j) is the vector whose i^{th} (resp. j^{th}) component equals 1, all others being null.

➠ Solution $y = (\mathbf{9}, 10, \mathbf{9}, 10)$ should be preferred to $x = (\mathbf{11}, 10, \mathbf{7}, 10)$ because there exists a transfer of size $\epsilon = 2$ (i.e. $\frac{11-7}{2}$) to pass from x to y.

Ordered Weighted Average (OWA)

- OWA is a family of aggregation functions which assigns importance weights to the **sorted values** of the utility function [Yager 88] :

$$W(x) = \sum_{k=1}^{n} w_i x_{\sigma(i)}$$

with $w = (w_1, \dots, w_n) \in [0,1]^n$ and $x_{\sigma(1)} \leq x_{\sigma(2)} \leq \cdots \leq x_{\sigma(n)}$

- **Proposition 1.** Schur-convex functions (also called order-preserving functions) are equitable aggregates :

$$x \succsim_{\|} y \Leftrightarrow \psi(x) \geq \psi(y)$$

- **Theorem 1**. Let be the following coefficients of the OWA aggregation: $W(x) = \sum_{k=1}^{n} sin(\frac{(n+1-k)\pi}{2n+1}) x_{(k)}$. W is a Schur-convex function.

➠ This result is fundamental, since Schur-convex functions ensure equity.

Basic ILP model for equitable conceptual clustering

- Each agent represents a concept and has its own utility corresponding to its frequency

$$
\max \sum_{c=1}^{|\mathcal{C}|} \omega_c \,.\, r_c
$$

$$
\text{s.t.} \begin{cases}
\text{Clustering.} \begin{cases} (C1) & \sum_{c=1}^{|\mathcal{C}|} a_{t,c} \,.\, x_c = 1 & \forall t \in \mathcal{T} \\ (C2) & k_{min} \leq \sum_{c=1}^{|\mathcal{C}|} x_c \leq k_{max} & \end{cases} \\
\texttt{OWA} \text{ sorting.} \begin{cases} (O1) & r_c - (v_i \,.\, x_i) \leq M \times z_{c,i} & \forall\ i, c = 1, ..., |\mathcal{C}| \\ (O2) & \sum_{i=1}^{|\mathcal{C}|} z_{c,i} \leq c - 1 & \forall\ c = 1, ..., |\mathcal{C}| \end{cases} \\
x_c \in \{0,1\},\ r_c \in R^+, \quad \forall\ c = 1, ..., |\mathcal{C}| \\
z_{c,i} \in \{0,1\}, \quad \forall\ i, c = 1, ..., |\mathcal{C}|
\end{cases}
$$

Basic ILP model for equitable conceptual clustering

- Each agent represents a concept and has its own utility corresponding to its frequency

$$
\begin{aligned}
&\texttt{max} \ \sum_{c=1}^{|\mathcal{C}|} \omega_c \,.\, r_c \\
&\text{s.t.} \begin{cases}
\text{Clustering.} \begin{cases}
\text{(C1)} & \sum_{c=1}^{|\mathcal{C}|} a_{t,c} \,.\, x_c = 1 & \forall t \in \mathcal{T} \\
\text{(C2)} & k_{min} \leq \sum_{c=1}^{|\mathcal{C}|} x_c \leq k_{max} &
\end{cases} \\
\texttt{OWA} \text{ sorting.} \begin{cases}
\text{(O1)} & r_c - (v_i \,.\, x_i) \leq M \times z_{c,i} & \forall\ i, c = 1, ..., |\mathcal{C}| \\
\text{(O2)} & \sum_{i=1}^{|\mathcal{C}|} z_{c,i} \leq c - 1 & \forall\ c = 1, ..., |\mathcal{C}|
\end{cases} \\
x_c \in \{0,1\},\ r_c \in R^+, \quad \forall\ c = 1, ..., |\mathcal{C}| \\
z_{c,i} \in \{0,1\}, \quad \forall\ i, c = 1, ..., |\mathcal{C}|
\end{cases}
\end{aligned}
$$

➠ requires $(|\mathcal{T}| + |\mathcal{C}|^2 + |\mathcal{C}| + 2)$ constraints and $(2 \times |\mathcal{C}| + |\mathcal{C}|^2)$ variables

Improved ILP model for equitable conceptual clustering

$$\max \sum_{c=1}^{|\mathcal{C}|} \omega_c \,.\, (v_c^{\uparrow} \,.\, x_c^{\uparrow})$$

$$\text{s.t.} \begin{cases} (\text{C1}), \quad (\text{C2}) \\ x_c \in \{0,1\}, \\ \quad \forall\ c = 1, ..., |\mathcal{C}| \end{cases}$$

- **Sorting constraints.** The utility values are known beforehand. Sorting is **performed immediately after** finding closed patterns.

Improved ILP model for equitable conceptual clustering

$$\texttt{max} \ \sum_{c=1}^{|\mathcal{C}|} \omega_c \,.\, (v_c^{\uparrow} \,.\, x_c^{\uparrow})$$

$$\text{s.t.} \begin{cases} (\text{C1}), \quad (\text{C2}) \\ x_c \in \{0,1\}, \\ \quad \forall\ c = 1, ..., |\mathcal{C}| \end{cases}$$

- **Sorting constraints.** The utility values are known beforehand. Sorting is **performed immediately after** finding closed patterns.

➠ requires $(|\mathcal{T}| + 2)$ constraints and $(|\mathcal{C}|)$ variables

Experiments – Evaluation

Tools for our 2-step approach

- Mining closed itemsets with `LCM` algorithm (without minimal frequency)
- Solving the ILP with `Cplex` version 12.4

Experimental evaluation

- Runtime and Scalability:
  - ILP models: `maxSum`, `maxMin`, `minDev` and `OWA`
  - CP models [Chabert et al., 2017] : `FullCP2` and `HybridCP` with `maxMin`

- Quality of balancing:
  - ILP models : different metrics
    1. ratio between the frequency of the smallest cluster to the average cluster frequency (i.e. Min/Avg)
    2. standard deviation in cluster frequencies (i.e. StdDev)
    3. deviation between the smallest and the largest description size (i.e. devSize)

Experiments – Datasets

Dataset	$\#\mathcal{T}$	$\#\mathcal{I}$	Density(%)	$\#\mathcal{C}$
Soybean	630	50	32	31,759
Primary-tumor	336	31	48	87,230
Lymph	148	68	40	154,220
Vote	435	48	33	227,031
tic-tac-toe	958	27	33	42,711
Mushroom	8124	119	18	221,524
Zoo-1	101	36	44	4,567
Hepatitis	137	68	50	3,788,341
Anneal	812	93	45	1,805,193

(a) UCI datasets.

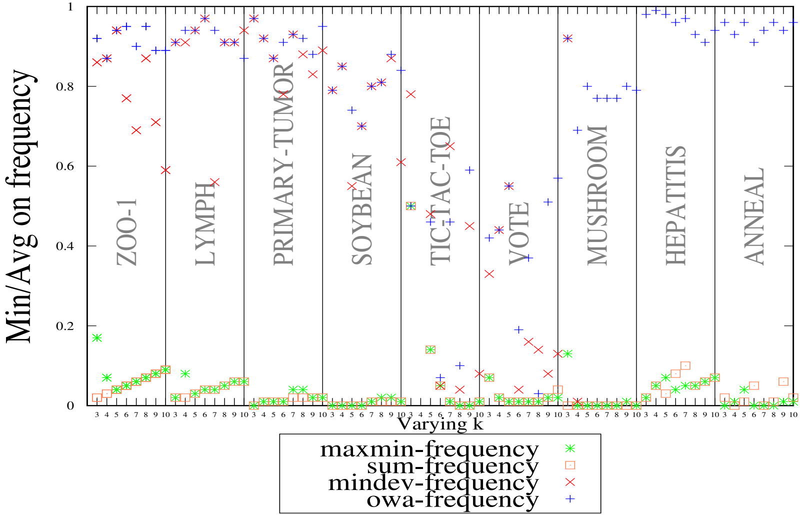
Dataset	$\#\mathcal{T}$	$\#\mathcal{I}$	Density(%)	$\#\mathcal{C}$
ERP-1	50	27	48	1,580
ERP-2	47	47	58	8,1337
ERP-3	75	36	51	10,835
ERP-4	84	42	45	14,305
ERP-5	94	53	51	63,633
ERP-6	95	61	48	71,918
ERP-7	160	66	45	728,537

(b) ERP datasets.

Table : Dataset characteristics. Each row gives the number of transactions ($\#\mathcal{T}$), the number of items ($\#\mathcal{I}$), the density and the number of closed patterns extracted ($\#\mathcal{C}$).

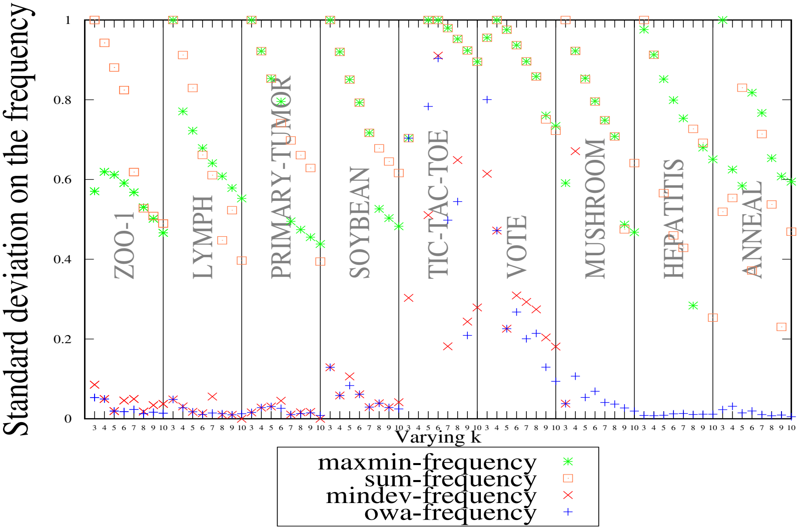
Results – Quality of balancing (1/2)

Min/Agv metric on frequency



Results – Quality of balancing (2/2)

Standard deviation metric on frequency



Runtime

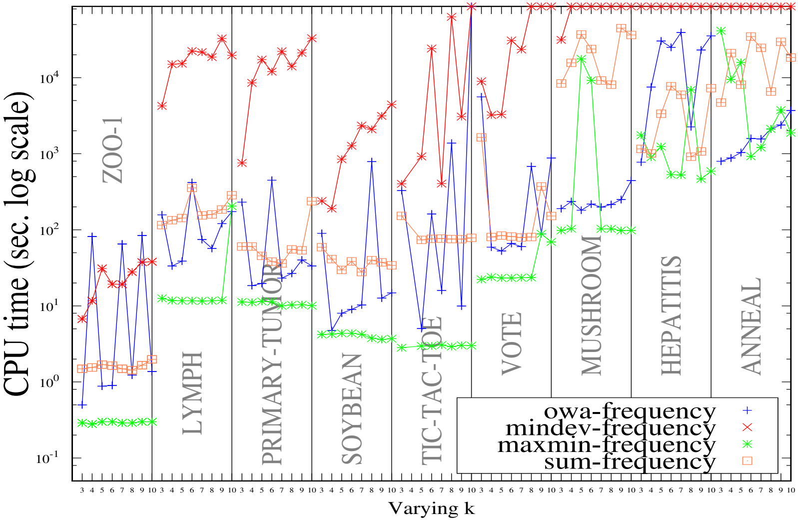


Figure : Comparing CPU-times of `maxMin` ILP models.

Conclusions

- OWA operator for implementing equity

- Optimal balanced conceptual clustering

- Scaling on larger datasets

Towards equitable multi-criteria conceptual clustering

Four conceptual clusterings for k=4

Sol.	X_1	X_2	X_3	X_4
s_1	{A} ($t_1..t_4$)	{B, E, G} (t_5, t_6)	{C, G} ($t_7..t_{11}$)	{C, E, H} (t_9, t_{10})
s_2	{B} ($t_1..t_6$)	{A, E} ($t_2..t_4$)	{C, G} ($t_7..t_{11}$)	{C, E, H} (t_9, t_{10})
s_3	{A, B, D} (t_1)	{A, E, F} (t_2)	{E, G} ($t_3..t_8$)	{C, H} ($t_9..t_{11}$)
s_4	{A, B, D} (t_1)	{F} (t_2, t_{11})	{E, G} ($t_3..t_8$)	{C, E, H} (t_9, t_{10})

- La solution $s3$ est plus équilibrée selon le critère **taille**
- La solution $s2$ est plus équilibrée selon le critère **fréquence**
- La solution $s1$ est plus équilibrée sur les deux critères **fréquence** et **taille**

Thanks for your attention !

Q & A