Clustering and identification of core implications

ICFCA 2021

D. López-Rodríguez[1], P. Cordero[1], M. Enciso[2], Á. Mora[1]

[1]Dep. de Matemática Aplicada
[2]Dep. de Lenguajes y Ciencias de la Computación

UNIVERSIDAD
DE MÁLAGA

Table of Contents

Introduction

Objectives of this work:

- To present briefly the `fcaR` package and the methods implemented.
- To present the research line on clustering of implications with an application example

The `fcaR` library

- `fcaR` is the first R package that implements the core notions and methods of FCA.
- It is designed to work with binary and fuzzy (graded) formal contexts.
- It is publicly available at the CRAN repository[1] and, up to June 2021, it has more than 10K downloads.

[1] `https://cran.r-project.org/package=fcaR`

Formal Contexts

```
fc <- FormalContext$new(planets)
fc$to_latex(caption = "The planets formal context")
```

	small	medium	large	near	far	moon	no_moon
Mercury	×			×			×
Venus	×			×			×
Earth	×			×		×	
Mars	×			×		×	
Jupiter			×		×	×	
Saturn			×		×	×	
Uranus		×			×	×	
Neptune		×			×	×	
Pluto	×				×	×	

Table 1: The planets formal context

```r
fc$clarify()
```

	small	medium	large	near	far	moon	no_moon
Pluto	×				×	×	
[Mercury, Venus]	×			×			×
[Earth, Mars]	×			×		×	
[Jupiter, Saturn]			×		×	×	
[Uranus, Neptune]		×			×	×	

Table 2: The clarified formal context

```r
S <- Set$new(fc$attributes, no_moon = 1)
fc$closure(S)
```

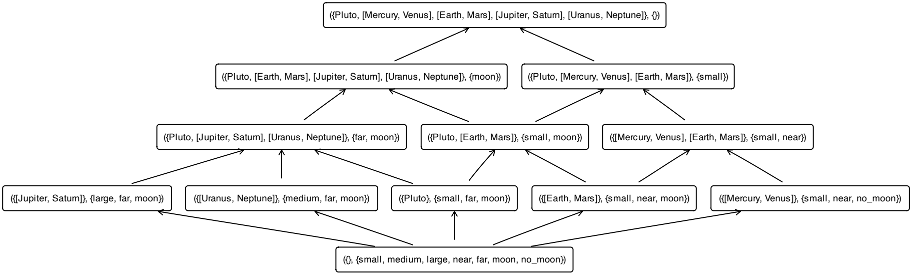
{small, near, no_moon}

```r
fc$att_concept("far")
```

({Pluto, [Jupiter, Saturn], [Uranus, Neptune]}, {far, moon})

Concept lattices

```
fc$find_concepts()
fc$concepts$plot()
```



```
fc$concepts$meet_irreducibles()
```

1: $(\{\text{Pluto}, [\text{Earth}, \text{Mars}], [\text{Jupiter}, \text{Saturn}], [\text{Uranus}, \text{Neptune}]\}, \quad \{\text{moon}\})$
2: $(\{\text{Pluto}, [\text{Jupiter}, \text{Saturn}], [\text{Uranus}, \text{Neptune}]\}, \quad \{\text{far}, \text{moon}\})$
3: $(\{[\text{Jupiter}, \text{Saturn}]\}, \quad \{\text{large}, \text{far}, \text{moon}\})$
4: $(\{[\text{Uranus}, \text{Neptune}]\}, \quad \{\text{medium}, \text{far}, \text{moon}\})$
5: $(\{\text{Pluto}, [\text{Mercury}, \text{Venus}], [\text{Earth}, \text{Mars}]\}, \quad \{\text{small}\})$
6: $(\{[\text{Mercury}, \text{Venus}], [\text{Earth}, \text{Mars}]\}, \quad \{\text{small}, \text{near}\})$
7: $(\{[\text{Mercury}, \text{Venus}]\}, \quad \{\text{small}, \text{near}, \text{no\_moon}\})$

- Computation of the (fuzzy) lattice Lattice operations: meet- and join-irreducible elements, infimum and supremum, sublattices, support…

Implications

```
fc$find_implications()
fc$implications
```

$$
\begin{array}{rrcl}
1: & \{\text{no\_moon}\} & \Rightarrow & \{\text{small, near}\} \\
2: & \{\text{far}\} & \Rightarrow & \{\text{moon}\} \\
3: & \{\text{near}\} & \Rightarrow & \{\text{small}\} \\
4: & \{\text{large}\} & \Rightarrow & \{\text{far, moon}\} \\
5: & \{\text{medium}\} & \Rightarrow & \{\text{far, moon}\} \\
6: & \{\text{medium, large, far, moon}\} & \Rightarrow & \{\text{small, near, no\_moon}\} \\
7: & \{\text{small, near, moon, no\_moon}\} & \Rightarrow & \{\text{medium, large, far}\} \\
8: & \{\text{small, near, far, moon}\} & \Rightarrow & \{\text{medium, large, no\_moon}\} \\
9: & \{\text{small, large, far, moon}\} & \Rightarrow & \{\text{medium, near, no\_moon}\} \\
10: & \{\text{small, medium, far, moon}\} & \Rightarrow & \{\text{large, near, no\_moon}\}
\end{array}
$$

```r
fc$implications$apply_rules(c("simplification", "rsimplification"))
fc$implications
```

$$
\begin{array}{rrcl}
1: & \{\text{no\_moon}\} & \Rightarrow & \{\text{near}\} \\
2: & \{\text{far}\} & \Rightarrow & \{\text{moon}\} \\
3: & \{\text{near}\} & \Rightarrow & \{\text{small}\} \\
4: & \{\text{large}\} & \Rightarrow & \{\text{far}\} \\
5: & \{\text{medium}\} & \Rightarrow & \{\text{far}\} \\
6: & \{\text{medium, large}\} & \Rightarrow & \{\text{no\_moon}\} \\
7: & \{\text{moon, no\_moon}\} & \Rightarrow & \{\text{medium, large}\} \\
8: & \{\text{near, far}\} & \Rightarrow & \{\text{medium, large}\} \\
9: & \{\text{small, large}\} & \Rightarrow & \{\text{medium}\} \\
10: & \{\text{small, medium}\} & \Rightarrow & \{\text{large}\}
\end{array}
$$

```r
S <- Set$new(fc$attributes, large = 1)
fc$implications$closure(S, reduce = TRUE)
```

```
$closure
{large, far, moon}

$implications
Implication set with 4 implications.
Rule 1: {no_moon} -> {medium, near}
Rule 2: {medium} -> {no_moon}
Rule 3: {near} -> {small, medium}
Rule 4: {small} -> {medium}
```

Clustering of implications

- We aim to study the potential use and applications of performing (unsupervised) clustering on the Duquenne-Guigues basis of implications.
- We show this idea using a running example.
- We try to get more insight with a more complex problem, to study the consistency of the clusters.

Idea

- Let us consider a formal context $\mathbb{K} = (G, M, I)$ and let Γ be the corresponding Duquenne-Guigues basis of implications.
- Find a partition $\Gamma = \Gamma_1 \cup \Gamma_2 \cup \ldots \cup \Gamma_K$ such that the quantity

$$\phi(\Gamma_1, \ldots, \Gamma_K) = \sum_{i=1}^{K} \delta(\Gamma_i)$$

  is minimum, where $\delta(\Gamma_i)$ represents an internal dissimilarity measure in Γ_i.
- Define a distance function between implications ($P \to Q$ and $R \to T$):
  - Appeareance (similarity between P and R, or Q and T)
  - Semantics (similarity between P^+ and R^+)
- Our intuition is that the pseudo-intents and the closed sets play an essential role in the clusters, but we want to explore the possibilities.

Let us suppose $A, B \subset M$. The following measures are based on well-known distances:

- Hamming (or Manhattan) distance: $d_{\mathrm{M}}(A, B) = |A \triangle B|$ (where $\triangle$ denotes the symmetric set difference operator) measures the amount of attributes that are present in only one of A and B.
- Jaccard index: $d_{\mathrm{J}}(A, B) = 1 - \dfrac{|A \cap B|}{|A \cup B|}$ measures the proportion of common attributes in A and B.
- Cosine distance: $d_{\cos}(A, B) = 1 - \dfrac{|A \cap B|}{\sqrt{|A| \cdot |B|}}$.

- Dissimilarity $\mathrm{dis}(P \to Q, R \to T)$ between two implications $P \to Q$ and $R \to T$

$$
\begin{aligned}
\mathrm{dis}_1(P \to Q, R \to T) &:= d(P, R) \\
\mathrm{dis}_2(P \to Q, R \to T) &:= d(P^+, R^+) \\
\mathrm{dis}_3(P \to Q, R \to T) &:= d(P, R) + d(Q, T) \\
\mathrm{dis}_4(P \to Q, R \to T) &:= d(P, R) + d(P^+, R^+) \\
\mathrm{dis}_5(P \to Q, R \to T) &:= d(P, R) + d(Q, T) + d(P^+, R^+)
\end{aligned}
$$

- The internal dissimilarity in the cluster:

$$
\delta(\Gamma_i) := \frac{1}{|\Gamma_i|} \sum_{R \to T \in \Gamma_i} \mathrm{dis}(P \to Q, R \to T)
$$

- We can compute a central implication in each cluster (the one that minimizes its distance to the other implications in the same cluster)
- Clustering algorithm: Partitioning Around Medoids (PAM), gives the central implications with the same cost
- Maybe necessary to remove implications with 0-support (since $P^+ = M$).

An example

Applying this strategy to the `planets` formal context, with this dissimilarity function:

$$\mathrm{dis}(P \to Q, R \to T) := |P \triangle R| + |P^+ \triangle R^+|$$

we obtain the following core implications:

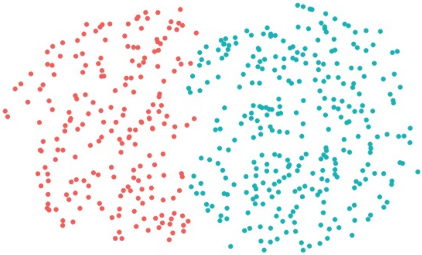
```
Implication set with 2 implications.
Rule 1: {near} -> {small}
Rule 2: {far} -> {moon}
# The cluster 1 is:
Implication set with 2 implications.
Rule 1: {no_moon} -> {small, near}
Rule 2: {near} -> {small}
# The cluster 2 is:
Implication set with 3 implications.
Rule 1: {far} -> {moon}
Rule 2: {large} -> {far, moon}
Rule 3: {medium} -> {far, moon}
```

The dataset

- We apply our proposal to the data from the so-called MONK's problems, a set of 3 datasets used in machine learning competitions.
- Each of the 3 datasets consists of 6 categorical attributes, `a1` to `a6`, taking integer values, and a binary `class` attribute.
- For this work, all categorical variables have been binarized, making an aggregate of 19 binary attributes, including the two class attributes, `class = 0` and `class = 1`.
- For each of these three problems, we have computed the Duquenne-Guigues basis, consisting of 524, 723 and 489 implications, respectively.
- After removing the implications that incorporate all the attributes, as commented before, the final sets of implications consisted of 505, 704 and 471 implications for problems MONKS-1, MONKS-2 and MONKS-3, respectively.

Clustering computation

- Silhouette index has determined that there are two clusters in each of the problems



- In MONKS-1, the core implications found were:

$$\{a5 = 1\} \Rightarrow \{class = 1\}$$
$$\{class = 0,\ a2 = 1,\ a5 = 2\} \Rightarrow \{a6 = 2\}$$

Clustering consistency

- Closure purity: Let us consider the set of *equivalence classes* in the Duquenne-Guigues basis Γ, as

$$[P \to Q] = \{R \to T \in \Gamma : P^+ = R^+\}$$

Table 3: Percentage of equivalence classes that belong to only one cluster

Problem	Dissimilarity	Hamming	Jaccard	Cosine
MONKS-1	dis_1	0.953	1.000	0.983
	dis_2	1.000	1.000	1.000
	dis_3	0.962	0.953	0.953
	dis_4	1.000	1.000	0.971
	dis_5	1.000	0.988	0.962
MONKS-2	dis_1	0.928	0.966	0.942
	dis_2	1.000	1.000	1.000
	dis_3	0.986	0.966	0.974
	dis_4	0.994	1.000	0.998
	dis_5	0.996	0.954	0.974
MONKS-3	dis_1	0.923	1.000	0.972
	dis_2	1.000	1.000	1.000
	dis_3	0.935	0.985	0.978
	dis_4	1.000	0.997	0.994
	dis_5	1.000	0.994	0.966

- Common attributes per cluster

Table 4: Attributes that appear more than 80% of the implications in each cluster.

		Hamming		Jaccard		Cosine	
Pr.	Diss.	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
1	dis_1	$\{class = 1\}$	$\varnothing$	$\{class = 1, a5 = 1\}$	$\varnothing$	$\{class = 1, a5 = 1\}$	$\varnothing$
	dis_2	$\{class = 1\}$	$\{class = 0\}$	$\{class = 1, a5 = 1\}$	$\varnothing$	$\{class = 1, a5 = 1\}$	$\varnothing$
	dis_3	$\{class = 1\}$	$\varnothing$	$\{class = 1\}$	$\varnothing$	$\{class = 1\}$	$\varnothing$
	dis_4	$\{class = 1\}$	$\{class = 0\}$	$\{class = 1, a5 = 1\}$	$\varnothing$	$\{class = 1\}$	$\varnothing$
	dis_5	$\{class = 1\}$	$\{class = 0\}$	$\{class = 1\}$	$\varnothing$	$\{class = 1\}$	$\varnothing$
2	dis_1	$\varnothing$	$\varnothing$	$\{a5 = 1\}$	$\varnothing$	$\{a4 = 1\}$	$\varnothing$
	dis_2	$\{class = 0\}$	$\{class = 1\}$	$\{class = 0, a6 = 1\}$	$\varnothing$	$\{class = 0, a5 = 1, a6 = 1\}$	$\varnothing$
	dis_3	$\{class = 0\}$	$\varnothing$	$\{class = 0\}$	$\varnothing$	$\{class = 0\}$	$\varnothing$
	dis_4	$\{class = 0\}$	$\{class = 1\}$	$\{class = 0, a4 = 1, a6 = 1\}$	$\varnothing$	$\{class = 0\}$	$\varnothing$
	dis_5	$\{class = 0\}$	$\varnothing$	$\{class = 0\}$	$\varnothing$	$\{class = 0\}$	$\varnothing$
3	dis_1	$\varnothing$	$\{class = 1\}$	$\{class = 0, a5 = 4\}$	$\varnothing$	$\{class = 0, a5 = 4\}$	$\varnothing$
	dis_2	$\{class = 0\}$	$\{class = 1\}$	$\{class = 0, a5 = 4\}$	$\varnothing$	$\{class = 0, a5 = 4\}$	$\varnothing$
	dis_3	$\{class = 0\}$	$\varnothing$	$\{class = 0\}$	$\varnothing$	$\{class = 0\}$	$\varnothing$
	dis_4	$\{class = 0\}$	$\{class = 1\}$	$\{class = 0\}$	$\varnothing$	$\{class = 0\}$	$\varnothing$
	dis_5	$\{class = 0\}$	$\{class = 1\}$	$\{class = 0\}$	$\varnothing$	$\{class = 0\}$	$\varnothing$

Conclusions and future work

- We have presented the `fcaR` package developed in the R language.
  - It provides a tool to the FCA community to test and compare algorithms and ideas.
  - It aims at making FCA works visible to other areas as machine learning, data science, etc., where the use of the R language is widely extended.
  - The code and experiments are in `https://github.com/Malaga-FCA-group/FCA-ImplicationClustering`
- We propose a method to cluster implications
  - Extracting interesting knowledge about the central implications
  - New interesting research in current areas of interest as Social Network Analysis: the identification of topics could be addressed by our clustering implication method based on logic.

Future research

- Natural clusters (consistent with the data) seem to emerge from the implication clusters, and this could have potential applications:
  - To reduce the concept lattice,
  - To simplify the bases of implications,
  - To research new algorithms to compute approximate closures.
- Study the relationship between the concept lattice obtained directly from a formal context and obtained after clustering objects.
- The study of *closure purity* can reveal interesting properties about closed sets and their features.
- Key attributes arise from the clusters, with potential applications revealing attributes and object clusters and their leaders.

Clustering and identification of core implications

ICFCA 2021

D. López-Rodríguez[1], P. Cordero[1], M. Enciso[2], Á. Mora[1]

[1]Dep. de Matemática Aplicada
[2]Dep. de Lenguajes y Ciencias de la Computación

UNIVERSIDAD DE MÁLAGA