Geometric Formal Concept Analysis

1. Abstract data cube category and lattice. Meet and join.
2. Binding preliminaries. Maximal clusters.
3. Concrete data cube and description diagrams. Feature matrix, signatures, formal concepts.
4. The Formal Concept Analysis approach. Adjoint functors, Galois connection, closure, concept lattice.
5. Multiple concept binding. Tetrahedral and cubical points of view. Basic concepts.
6. Evaluation and selection of optimal description diagrams. Iterative binding algorithms.
7. Visualization of the final description. Decision tree row/column orders. Concept quality.
8. Further analysis of a description. Syntactic and semantic transfer.

Geometric Formal Concept Analysis

① Abstract data cube category and lattice

$C = [0,1]^n$ data cube

$p_1, \ldots, p_n$ coordinate functions on C; properties

$C_n := \{C\}$

$C_{n-1} := \{Z(p_i - v) \mid i=1,\ldots,n \quad v=0,1\}$ generator facets; faces

$C_k := \{\bigcap_{f \in G} f \mid G \subset C_{n-1},\ |G| = n-k\} \setminus \{\emptyset\}$ k-dimensional facets

$\mathcal{C} := (\coprod_{k=0}^{k=n} C_k,$ set inclusions$)$ facet category

$\psi : \mathcal{P}(C_{n-1}) \to \mathcal{P}(C) \quad G \mapsto \cap G$ geometric realization of symbolic facets

$A = \{G \in \mathcal{P}(C_{n-1}) \mid \psi G \neq \emptyset\}$

$\psi|_A$ is injective.

$\mathrm{Obj}\,\mathcal{C} = \mathrm{Image}\,\psi \setminus \{\emptyset\}$

$\vee := (\psi|_A)_*(\cup)$ meet in $\mathcal{C}$

$\wedge := (\psi|_A)_*(\cap)$ join in $\mathcal{C}$

$\mathcal{C}$ is a poset category.

In $\mathcal{C}$ limits and colimits depend only on diagram objects; are products and coproducts.

$F \subset \mathrm{Obj}\,\mathcal{C}$

<u>Proposition</u> $\lim F = \vee F \qquad \mathrm{colim}\, F = \wedge F$
$\qquad = \sup F \qquad\qquad = \inf F$

$\left.\begin{array}{l}\mathrm{Obj}\,\mathcal{C} \to C \quad f \mapsto \text{midpoint } f \\ \mathrm{Mor}\,\mathcal{C} \to \text{Line segments in } C\end{array}\right\}$ geometric realization of $\mathcal{C}$ as lattice

② Binding preliminaries (compare [E.V., MES] "cluster" and "complex link")

$\tilde{\mathcal{C}}$ arbitrary category

$\mathrm{Diag}\,\tilde{\mathcal{C}}$ diagrams in $\tilde{\mathcal{C}}$

$D_1, D_2 \in \mathrm{Diag}\,\tilde{\mathcal{C}}$

$\{m_{ab} : a \to b\}_{a \in D_1,\, b \in D_2}$ a system of morphisms

$\{m_{ab}\}$ called a <u>cluster</u> if commute with D_1, D_2

cluster $\{m_{ab}\}$ called <u>simple</u> if factor through some $a' \to b'$.

<u>Proposition</u> Assume $\tilde{\mathcal{C}}$ is a poset category. Let $E, F \subset \mathrm{Ob}_j\,\tilde{\mathcal{C}}$
There is a cluster $E \to F$ if and only if $\operatorname{colim} E \to \lim F$.
It is automatically simple, and unique.

cluster $E \to F$ called <u>maximal</u> if $\operatorname{colim} E \xrightarrow{\cong} \lim F$

③ Concrete data cube and description diagrams

M — $N \times n$ binary matrix table; observations or samples; feature matrix

S — set of N rows of M; signatures (assume non-redundant)

$S \subset C_0$

Assume $\wedge S = C$. Otherwise reduce from C to $\wedge S$ by removing redundant properties.

Draw diagram in $\mathcal{C}$, one morphism $s \to (p,v)$ $(s \in S,\ (p,v) \in C_{n-1})$ for each entry of M.

D — diagram in $\mathcal{C}$

D called a description (of S) if the subcategory of $\mathcal{C}$ generated by D contains Draw

A (formal) concept, in the context of M or S, is a maximal cluster $E \to F$, $E \subset S$ and $F \subset C_{n-1}$.

Descriptions can be generated by binding concepts.

④ The Formal Concept Analysis approach

$L := \psi^{-1} \circ \wedge : \mathcal{P}(S) \to \mathcal{P}(C_{n-1})$

$R := V(-) \cap S : \mathcal{P}(C_{n-1}) \to \mathcal{P}(S)$

L, R are adjoint functors, yielding Galois connection and closure operators:

$$E \mapsto \overline{E} = R(L(E)) \qquad E \subset S$$

$$F \mapsto \overline{F} = L(R(F)) \qquad F \subset C_{n-1}$$

theorem (Rudolph Wille) The concepts form a complete lattice. Projection to domain E or codomain F yields an isomorphism to lattice of closed subsets of S and closed subsets of C_{n-1} respectively. (The latter being contravariant).

theorem/algorithm (   ) The closed sets (and hence the concepts) can be enumerated by recursive closure of pairwise unions, starting from singletons in S.

theorem/algorithm (   ) The closed sets can be enumerated in "lectic order" with respect to ordering of $\{p_1, \ldots, p_n\}$

The FCA approach is concerned with just one description of S: the concept lattice.

⑤ Multiple concept binding; tetrahedral and cubical points of view

$e := |E| \quad f := |F|$

Binding $E \to F$ changes #edges: $ef \to e+f$, #nodes: $+1$ (when starting from D_{raw})

If two concepts are completely independent, their edge/node savings are also independent (they add).

$E \to F$, $E' \to F'$ two concepts (say, c and c')

$e_1 := |E \setminus E'| \quad e_2 := |E \cap E'| \quad e_3 := |E' \setminus E|$

$f_1 := |F \setminus F'| \quad f_2 := |F \cap F'| \quad f_3 := |F' \setminus F|$

$e := e_1 + e_2 + e_3 = |E \cup E'|$

$f := f_1 + f_2 + f_3 = |F \cup F'|$

$u := E \cup E' \to F \cap F'$ upper concept induced by c, c'

$l := E \cap E' \to F \cup F'$ lower concept induced by c, c'

Binding c, c', u, l changes #edges: $\underbrace{\sum_{i=1}^{3} e_i f_i + \sum_{j=1}^{2} (e_j f_{j+1} + e_{j+1} f_j)}_{ef - (e_1 f_3 + e_3 f_1)} \to e + f + 4$, #nodes: $+4$

General case, K concepts $c_1, \dots, c_K$

If all intersections are non-empty, the concepts induced by $\{c_i\}$ are labelled by the vertices of the K-dimensional cube; equivalently, by the facets of the $(K-1)$-dimensional simplex. Aside from morphisms emanating directly from S or C_{n-1}, the morphisms of the bound diagram are labelled by the 1-skeleton of the K-cube.

If some intersections are empty, one needs the quotient graph by the edges emanating from the nodes labelling empty intersection.

<u>Problem</u> Identify appropriately minimal collection of concepts which induce the rest.

Attempted solution.

$E \to F$ called <u>basic</u> if one of the following equivalent conditions hold:

- E is neither the union of two proper closed subsets nor the intersection of two proper closed supersets.
- F is ....
- E is not the union of two proper closed subsets and F is not the union of two proper closed subsets.
- E is not the intersection of two proper closed supersets and F is not the intersection of two proper closed supersets.

<u>Problem</u> Identify all basic concepts with an algorithm.

⑥ Evaluation and selection of optimal description diagrams

Complexity metrics/functions on set of descriptions
- total number of edges
- total number of edges + total number of nodes
- total edge length in 2C embedding (Steiner problem)

~~Algorithmic strategies~~

Algorithm initialization strategies
- Sequence. Start from a deterministic pre-defined sequence of concepts to be bound, e.g. lectic order.
- Basis. Start from naive simultaneous binding of a pre-defined finite set of core concepts, e.g. basic concepts.

Bind until incremental improvement of complexity metric falls below error threshold, or until next step is not an improvement.

In basis case, should consider intersection/union of all pairs of previously bound concepts, or else a level-based method:

Level 1. Bind all pairs intersection/unions of basis until improvement threshold, replacing the pair with the new, intersection and union. (Still K concepts on the dynamic list).

Level 2. Bind all triples intersection/unions of basis until improvement threshold, replacing the 3 with the 2 new, intersection and union.

... termination is likely since the number of concepts on deck after Level 1 is decreasing with each increment.

⑦ Visualization of the final description

Append concept membership vectors to M as new features, use decision tree w.r.t. them to order columns of M.

Append concept membership vectors to M as quasi-samples, use decision tree w.r.t. them to order rows of M.

Use opacity map to display "quality" of description per entry of M, e.g.
- the size (area) of the largest concept capturing that entry
- average size (area) of the concepts capturing that entry
- as above, assessing concept quality differently, e.g. variance of feature values pre-dichotomization, in case M was created by dichotomization from continuous values.

(8) Further analysis of a description: Syntax/semantics and transfer

D a description of S

α a subdiagram of D containing C_{n-1} and none of S (α for "abstract")

K a subdiagram of D containing S and none of C_{n-1} (K for "concrete")

A splitting of D is α and K such that D is the union of α, K, and any morphisms sourced in K and targeted in α which belong to D.

Transfer to a new formal context M'/S' with at least the same feature set as M/S,

means the description D' of S' obtained by binding all closures of the upper sets of each object of α.

Transfer may be used within a single dataset to set up cross-validation.

The above may be called syntactic transfer.

A transfer procedure involving K may be called semantic transfer:

Extend to a new formal context M'/S' with more features on the same sample set S, e.g. outcome features previously withheld. Calculate D' by closure of each lower set of an object of K in the new context.