R. Belohlavek, M. Trnecka: Basic Level in Formal Concept Analysis: An Investigation in the Psychology of Concepts

Department of Computer Science, Palacky University, Olomouc (17. listopadu 12, CZ–77146 Olomouc, Czech Republic)

Introduction

We present a study regarding basic level of concepts in conceptual categorization. The basic level of concepts is an important phenomenon studied in the psychology of concepts. We propose to utilize this phenomenon in formal concept analysis to select important formal concepts. Such selection is critical because, as is well known, the number of all concepts extracted from data is usually large. We review and formalize the main existing psychological approaches to basic level which are presented only informally and are not related to any particular formal model of concepts in the psychological literature. Interestingly, our formalization and experiments reveal previously unknown relationships between the existing approaches to basic level. Thus, we argue that a formalization of basic level in the framework of formal concept analysis is beneficial for the psychological investigations themselves because it helps put them on a solid, formal ground.

Basic Level Phenomenon

- Extensively studied phenomenon in psychology of concepts.
- When people categorize (or name) objects, they prefer to use certain kind of concepts.
- Such concepts are called the concepts of the basic level.
- Definition of basic level concepts?:
  Are cognitively economic to use; "carve the world well".
- One feature: Basic level concepts are a compromise between the most general and most specific ones.
- Several informal definitions proposed.

Basic Level Formalization

- For a given approach M to basic level, we define a function BL_{M} mapping every concept $\langle A,B\rangle$ in the concept lattice $\mathcal{B}(X,Y,I)$ to $[0,\infty)$ or to $[0,1]$.
- $BL_{\mathrm{M}}(A,B)$ is interpreted as the degree to which $\langle A,B\rangle$ belongs to the basic level.
- A basic level is thus naturally seen as a graded (fuzzy) set rather than a clear-cut set of concepts.
- We consider the following probability space: X (objects) are the elementary events, 2^X (sets of objects) are the events, the probability distribution is given by $P(\{x\})=\frac{1}{|X|}$ for every object $x\in X$. For an event $A\subseteq X$ then, $P(A)=|A|/|X|$. The event corresponding to a set $\{y,\dots\}\subseteq Y$ of attributes is $\{y,\dots\}^{\downarrow}$.

Basic Level Metrics

Similarity approach (S)

- Basic level concept satisfies three conditions:
  1. The objects of this concept are similar to each other;
  2. The objects of the superordinate concepts are significantly less similar;
  3. The objects of the subordinate concepts are only slightly more similar.
- Formalized in our previous paper (Belohlavek, Trnecka, ICFCA 2012).

Cue Validity (CV)

- Based on the notion of a cue validity of attribute y for concept c, i.e. the conditional probability $p(c|y)$ that an object belongs to c given that it has y.

$$BL_{\mathrm{CV}}(A,B)=\sum_{y\in B}P(A|\{y\}^{\downarrow})=\sum_{y\in B}\frac{|A\cap\{y\}^{\downarrow}|}{|\{y\}^{\downarrow}|}$$

Category Feature Collocation Approach (CFC)

- Defined as product $p(c|y)\cdot p(y|c)$ of the cue validity $p(c|y)$ and the so-called category validity $p(y|c)$.

$$BL_{\mathrm{CFC}}(A,B)=\sum_{y\in Y}\left(\frac{|A\cap\{y\}^{\downarrow}|}{|\{y\}^{\downarrow}|}\cdot\frac{|A\cap\{y\}^{\downarrow}|}{|A|}\right)$$

Category Utility Approach (CU)

- Utilizes the notation of category utility

$$cu(c)=p(c)\cdot\sum_{y\in Y}[p(y|c)^2-p(y)^2].$$

$$BL_{\mathrm{CFC}}(A,B)=P(A)\cdot\sum_{y\in Y}\left[\left(\frac{P(\{y\}^{\downarrow}\cap A)}{P(A)}\right)^2-P(\{y\}^{\downarrow})^2\right]=$$
$$=\frac{|A|}{|X|}\cdot\sum_{y\in Y}\left[\left(\frac{|\{y\}^{\downarrow}\cap A|}{|A|}\right)^2-\frac{|\{y\}^{\downarrow}|}{|X|})^2\right]$$

Predictability Approach (P)

- Frequently formulated in the literature.
- Basic level concepts are abstract concepts that still make it possible to predict well the attributes of their objects.
- We introduce a graded (fuzzy) predicate $pred$ such that $pred(c)\in[0,1]$ is interpreted as the truth degree of proposition "concept $c=\langle A,B\rangle$ enables good prediction".
- We use the principles of fuzzy logic to obtain the truth degrees β_1, β_2, and β_3 of the following three propositions:
  1. c has high $pred$;
  2. c has a significantly higher $pred$ than its upper neighbors;
  3. c has only a slightly smaller $pred$ than its lower neighbors.
- Basic level metric: $BL_{\mathrm{P}}(A,B)=\beta_1\otimes\beta_2\otimes\beta_3$. $\otimes$ represent appropriate truth function of many-valued conjunction.
- For a given concept $c=\langle A,B\rangle$ and attribute $y\in Y$, consider the random variables $V_y: X\to\{0,1\}$ and $V_c: X\to\{0,1\}$ defined by:

  $V_y(x)=1$ if $\langle x,y\rangle\in I$ and $V_y(x)=0$ if $\langle x,y\rangle\notin I$, and
  $V_c(x)=1$ if $x\in A$ and $V_c(x)=0$ if $x\notin A$.

  The fact that the value of y is well predictable for objects in c corresponds to the fact that the conditional entropy $E(V_y|V_c=1)$ is low.

$$E(V_y|V_c=1)=-\frac{|A-\{y\}^{\downarrow}|}{|A|}\cdot\log\frac{|A-\{y\}^{\downarrow}|}{|A|}-\frac{|\{y\}^{\downarrow}\cap A|}{|A|}\cdot\log\frac{|\{y\}^{\downarrow}\cap A|}{|A|}$$

- Averaging over all the attributes in $Y-B$ (because for $y\in B$ we have $E(V_y|V_c=1)=0$), we get quantity:

$$p(c)=\sum_{y\in Y-B}\frac{E(V_y|V_c=1)}{|Y-B|}.$$

- Since a low value of p corresponds to a good ability to predict by c, i.e. to a high value of $pred(c)$, letting

$$pred(c)=1-p(c).$$

Comparison of Basic Level Metrics

Similarity of Rankings

- For input data $\langle X,Y,I\rangle$, a given metric BL_{M} determines a ranking of formal concepts in $\mathcal{B}(X,Y,I)$, i.e. determines the linear quasiorder $\leq_{\mathrm{M}}$ defined by

$$\langle A_1,B_1\rangle\leq_{\mathrm{M}}\langle A_2,B_2\rangle \text{ iff } BL_{\mathrm{M}}(A_1,B_1)\leq BL_{\mathrm{M}}(A_2,B_2).$$

- We examined the pairwise similarities of the rankings $\leq_{\mathrm{S}}$, $\leq_{\mathrm{CV}}$, $\leq_{\mathrm{CU}}$, $\leq_{\mathrm{CFC}}$, and $\leq_{\mathrm{P}}$ for various datasets. We used the Kendall tau coefficient to assess the similarities.

	S	CV	CFC	CU	P
S	1.000	-0.093	-0.215	-0.139	-0.175
CV	-0.093	1.000	**0.737**	**0.754**	0.170
CFC	-0.215	0.737	1.000	**0.789**	0.229
CU	-0.139	0.754	0.789	1.000	0.161
P	-0.175	0.170	0.229	0.161	1.000

Kendall tau coefficients for synthetic 75×25 datasets

Similarity of sets of top r basic level concepts

- One is arguably more interested in the set consisting of the top r concepts of $\mathcal{B}(X,Y,I)$ according to the ranking $\leq_{\mathrm{M}}$ for a given metric M. We denote such set by Top_r^{M}.
- For formal concepts $\langle C,D\rangle,\langle E,F\rangle\in\mathcal{B}(X,Y,I)$, denote by $s(\langle C,D\rangle,\langle E,F\rangle)$ degree of similarity. For two metrics M and N, and a given $r=1,2,3,\dots$, we define:

$$S(Top_r^{\mathrm{M}},Top_r^{\mathrm{N}})=\min(I_{\mathrm{MN}},I_{\mathrm{NM}})$$

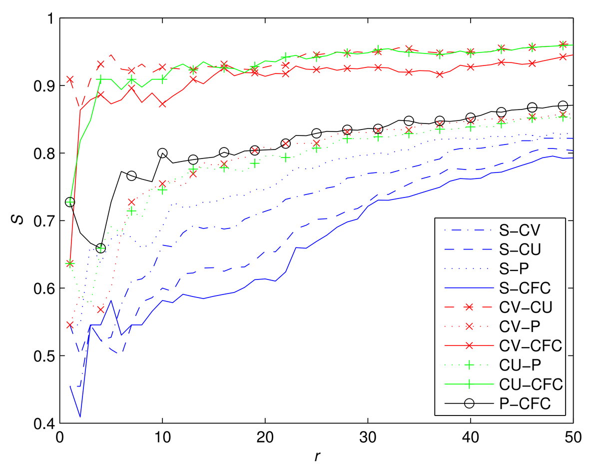
where

$$I_{\mathrm{MN}}=\frac{\sum_{\langle C,D\rangle\in Top_r^{\mathrm{M}}}\max_{\langle E,F\rangle\in Top_r^{\mathrm{N}}}s(\langle C,D\rangle,\langle E,F\rangle)}{|Top_r^{\mathrm{M}}|}$$

and

$$I_{\mathrm{NM}}=\frac{\sum_{\langle E,F\rangle\in Top_r^{\mathrm{N}}}\max_{\langle C,D\rangle\in Top_r^{\mathrm{M}}}s(\langle C,D\rangle,\langle E,F\rangle)}{|Top_r^{\mathrm{N}}|}.$$

- $S(Top_r^{\mathrm{M}},Top_r^{\mathrm{N}})$ may naturally be interpreted as the truth degree of the proposition "for most concepts in Top_r^{M} there is a similar concept in Top_r^{N} and vice versa".



Similarities S of sets of top r concepts for 75×25 random datasets.

Conclusions

CU, CFC, and CV may naturally be considered as a group of metrics with significantly similar behavior, while S and P represent separate, singleton groups. This observation contradicts the current psychological knowledge. Namely, the (informal) descriptions of S, P, and CU are traditionally considered as essentially equivalent descriptions of the notion of basic level in the psychological literature. On the other hand, CFC has been proposed by psychologists as a supposedly significant improvement of CV and the same can be said of CU versus CFC.

International Center for Information and Uncertainty
SSIU 2013: International Summer School "Information and Uncertainty"
http://mcin.upol.cz/SSWIU-2013/

european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
INVESTMENTS IN EDUCATION DEVELOPMENT