DATA FORMATION AND PROCESSING IN FORMAL CONCEPT ANALYSIS: SUBJECTIVE ASPECTS

D.E. Samoilov[1,2], S.V. Smirnov[1]

[1]Institute for the Control of Complex Systems, Russian Academy of Science,
[2]Samara National Research University, Samara, Russia

Abstract. The paper gives a brief overview of the subjective aspects of data formation and processing in Formal Concept Analysis. It is shown that the fundamental cognitive scaling procedure that allows a different interpretation, introduces new information into the analysis and the analysis is not correct in the general case without paying the proper attention to this information. The relationship between the objects properties that arises from the use of various types of scales and that need to be noted, is considered.

Introduction

For more than three decades the Formal Concept Analysis (FCA) is being developed successfully at the intersection of applied mathematics and computer science [1-7]. FCA has made a significant contribution and will continue to stimulate the developing of data mining, data representation and other parts of computer science due to the classical (Aristotelian) approach to the concept as the fundamental mental entity defined by the volume and content as well as to the basis of algebraic lattices theory.
FCA cognitive character appears in the account of the researcher's different axiological systems. The outline of FCA subjective aspects and its application in data analysis is the scope of this article. But the main focus is concentrated on primary data scaling. We believe that the genesis of so-called "properties' existence constraints" [8, 9], without which the FCA problems solution is incorrect [10], is often determined by scaling procedures [11, 12]. The occurrence of various restrictions of properties' existence constraints as a result of subjective selection of scales type is investigated.

1   Subjective aspects of classical FCA

1.1   Basic definitions and models

FCA has to deal with mass encountered practical applications that require the object-attributive data analysis. Classical FCA is focused on processing of binary data as a set of truth values of basic semantic proposition b_{gm} = "g object has m property". It uses the following symbols and models:

- $K = (G^*, M, I)$ – formal context where G^* is a set of investigated knowledge domain's objects (KD) comes in the researcher's view (i.e. the "learning sample" of KD objects), M – set of objects' measured properties, I – relation between the objects and their properties - a set of assessments $\|b_{gm}\| \in \{\textbf{True}, \textbf{False}\}$;

- Galois operators φ, ω (a common notation " ' ") for the context K:

  - $\varphi(X) = X' = \{m \mid m \in M, \forall g \in X \ ((g, m) \in I)\}$ - common objects' properties $X \subseteq G^*$;
  - $\omega(Y) = Y' = \{g \mid g \in G^*, \forall m \in Y \ ((g, m) \in I)\}$ – objects that have all the properties of the $Y \subseteq M$;
  - for a set of objects X, the set of their common properties X' is the description of the objects' similarity from the set X, and the closed set X'' is a cluster of similar objects;

- (X, Y) – formal concept where $X \subseteq G^*$ is extension, $Y \subseteq M$ is intention, $X = Y'$, $Y = X'$;

- $B(K)$ –set of all formal concepts of K;

- $(B(K), \leq)$ – concept's lattice where $(X_1, Y_1) \leq (X_2, Y_2)$, if $X_1 \subseteq X_2$ (or $Y_1 \supseteq Y_2$).

The subjective aspect of K context formation is manifested in cognitive asymmetry of "objects" and "properties": formally the objects G^* are independent from the researcher's KD, while the properties of M are the result of KD hypotheses production maid by the subject and it is based on his current target system, his a priori knowledge and his resource capabilities.

1.2   Formal concept's set reduction

The FCA results presentation for subsequent analysis may be difficult because of the large number of detectable concepts. Two main ways of relevant formal concepts' selection are developed for the reduction of the set $B(K)$.
The support for multiple properties $Y \subseteq M$ for a given context K is
$$\textbf{supp}(Y) = |Y'| / |G^*| .$$
The set $Y \subseteq M$ is called a frequent set of properties, if $\textbf{supp}(Y) \geq minsupp \in [0, 1]$.
If there are frequent concepts saved in the lattice only (their content is frequent sets of properties), the lattice will be reduced to the so-called "iceberg concepts" [13, 14].

The more granular approach is based on the identification in the $\boldsymbol{B}(\boldsymbol{K})$ the concepts that are resistant to the support volume's changes in the objects' learning sample [15, 16].

The stability index of the formal concept (X, Y) is determined by

$\sigma(X, Y) = |\{Z \subseteq X \mid Z' = Y\}| / |2^X|.$

The concept (X, Y) is considered to be stable when the $\sigma(X, Y) \geq \sigma_{min} \in [0, 1]$, and the lattice reducing means that the most stable formal concepts will be stored there only.

It is obvious that the subjective nature of the thresholds choice for the properties' variety support as well as for the concepts' sustainability index is not associated with the involvement of the additional information (knowledge) about the KD in the analysis.

1.3   Properties subsets' implications

The implication on formal context's properties subsets $\boldsymbol{K}$ is a dependence $A \rightarrow B$, $A, B \subseteq M$, provided that all objects with properties A, also have all the properties of B, i.e. $A' \subseteq B'$. Partial implication in the context $\boldsymbol{K}$ is distinguished by the lack of support in the objects' learning sample [1].

Entered into the FCA the partial implication's *confidence index* makes it possible to extend the set of relevant empirical regularities with condition of subjectively choosing of the reliability threshold. But it doesn't accompanied by using of additional data about KD.

2   Subjective aspects of conceptual scaling

The basic form of empirical information about KD is an "object-properties" table, which is treated in the FCA as a multi-valued context $(G^*, M, V, \boldsymbol{I})$. Here G^* and M have been already defined, V – is the property values' set, and $\boldsymbol{I}$ - is the ternary relation between G^*, M and V ($\boldsymbol{I} \subseteq G^* \times M \times V$) defined for all pairs from $G^* \times M$.

To reducing the many-valued context to a binary form, the conceptual scaling as a fundamental cognitive procedure is applied [1, 11]. It informally means the subjective construction of value domain's "coverage" of each property of multi-valued context, i.e., the formation of new KD objects' distinctive properties that are measured in subjectively formed scales.

The property scale $m \in M$ is a binary context $S_m = (G_m, M_m, I_m)$. Here G_m is the scale values, M_m – is the KD objects' new properties that are entered by the scale, I_m – is a relation between the scale values and the new properties introducing the specific of the KD subjective perception by its researcher.

We will show that subject enters qualitatively new information about KD into the analysis while it implements a conceptual scaling. FCA practical application becomes problematic without taking this information into account (these problems were discussed in [10]).

2.1   Using of the nominal scale

The most common scaling reception is the use of the nominal scale [11, 17]. Table 1 gives an example of such scale.

Table 1. Scale of men's growth

Growth, cm	Low	Average	High
< 168	×		
168-175		×	
> 175			×

The covering of the original values' domain of scaling property is strictly disjunctive in this case; items' fuzzy scale can be a model of a more complex approach to this problem.

It is obvious that E pair incompatibility [8, 10] (for example, E(Low, High)) is inherent by introduced nominal scale KD objects' properties in either embodiment. It is new essential information about KD that the researcher adds to the existing data in the original multi-valued context.

2.2   Other types of scales

Specific areas of conceptual explorations - such as sociology [18] or machine vision [19] - typically characterized by self-built complex types of scales.

We will show the effects of using of other types of scales on the examples from [20]. These examples do not embrace all of the possible methods of expression of the researcher's subjective perception of KD.

The ordinal scale should be used to preserve the values ordering in the domain of multi-valued property.

So, the domain of multi-valued properties named "Financial position" (FP) can be described by the following expressions (from "difficult" to "safe") [20]:

1. not enough money even for food;

2. enough money for food, but not enough to buy clothes and shoes;

3. have enough money for clothes and shoes, but can't afford the purchase of household appliances;

4. enough money to buy household appliances, but not enough to buy a new car;

5. enough money for everything, except the expensive acquisitions such as an apartment, a house;

6. do not feel financial difficulties, could buy an apartment, a house, etc., if necessary.

The researchers will have table 2 as the most natural scale for this multi-valued property.

Table 2. Financial position scale

	FP₁	FP₂	FP₃	FP₄	FP₅	FP₆
1	×					
2	×	×				
3	×	×	×			
4	×	×	×	×		
5	×	×	×	×	×	
6	×	×	×	×	×	×

This scaling sets the binary conditionality relation between newly introduced properties: C [8, 10]: $i < k \leftrightarrow C(\mathrm{FP}_k, \mathrm{FP}_i)$.

Nowadays, the scales with division and ordering become very popular. They are described in [20] to a closed question like "Do you feel safe?" (S). The response options are:

1. definitely yes;

2. rather yes;

3. rather no;

4. definitely no.

The subjective understanding of this domain of values can be expressed by double ordering scale (table 3).

Table 3. Safety scale

	S₁	S₂	S₃	S₄
1	×	×		
2		×		
3			×	
4			×	×

In this example, the researcher expands the available empirical evidence about KD subjectively by entering the following binary relations between the newly introduced properties:

- $E = \{(S_1, S_3), (S_1, S_4), (S_2, S_3), (S_2, S_4)\}$;
- $C = \{(S_1, S_2), (S_4, S_3)\}$.

Conclusion

Fundamental subjective aspect of the FCA is an axiological basis of formation of the initial data about KD. This aspect reveals itself in the formation of measurement procedures' set.

Subjectively established thresholds of different indicators are generally used to generate equivalence classes for the results and are directly interpreted in the terms of FCA. Fundamental cognitive scaling procedure, on the other hand, is associated with the subject's introduction of additional information about studied KD. This information should be taken into account at the stage of binary formal context formation [10] and it has a significant effect on derivable formal concepts' structure.

Of course, the genesis of the existence limits of the properties is not exhausted by the researcher's subjective actions during scales' designing for the property values of objects seen in the learning sample. Subject's a priori knowledge relevant to the researched KD is the source of these restrictions in general.

Acknowledgements

References

1. Ganter B, Wille R. Formal Concept Analysis. Mathematical foundations. Springer Berlin-Heidelberg, 1999.
2. Carpineto C, Romano G. Concept Data Analysis: Theory and Applications. Wiley, 2004.
3. Ganter B, Obiedkov S. Conceptual Exploration. Springer, 2016.
4. Begriffliche Wissensverarbeitung. Methoden und Anwendungen. Eds.: B. Ganter, R. Wille. Springer Berlin-Heidelberg, 2000.
5. Priss U, Szathmary L. Preface to the Special Issue on Concept Lattices and Their Applications - CLA 2012. Annals of Mathematics and Artificial Intelligence, 2014; 72(1): 1-2.
6. Ignatov DI. Introduction to Formal Concept Analysis and Its Applications in Information Retrieval and Related Fields. In: P. Braslavski, N Karpov, M. Worring, Y. Volkovich, D.I. Ignatov (Eds.): Information Retrieval (Revised Selected Papers 8th Russian Summer School, RuSSIR 2014. Nizhniy Novgorod, Russia, August 18-22, 2014). Springer International Publishing, 2015: 42-141.
7. Formal Concept Analysis Homepage. URL : http://www.upriss.org.uk/fca/fca.html.
8. Lammari N, Metais E. Building and maintaining ontologies: a set of algorithms. Data & Knowledge Engineering, 2004; 48(2): 155-176.
9. Pronina VA, Shipilina LB. Using the relationships between attributes to build domain ontology [In Russian]. Control Science, 2009; 1: 27-32.
10. Semenova VA, Smirnov SV. Intelligent analysis of incomplete data for building formal ontologies. CEUR Workshop Proceedings, 2016; 1638: 796-805. DOI: 10.18287/1613-0073-2016-1638-796-805
11. Ganter B, Wille R. Conceptual scaling. In: Applications of Combinatorics and Graph Theory to the Biological and Social Sciences. Ed.: F. Roberts. Springer Verlag, New York, 1989: 139-167.

12. Belohlavek R, Konecny J. Scaling, Granulation, and Fuzzy Attributes in Formal Concept Analysis. The IEEE International Conference on Fuzzy Systems (London, UK, July 23-26, 2007): 918-923.

13. Stumme G, Taouil R, Bastide Y, Pasqier N, Lakhal L. Computing Iceberg Concept Lattices with Titanic. Journal on Knowledge and Data Engineering, 2002; 42(2): 189-222.

14. Nehme K, Valtchev P, Rouane MH, Godin R. On Computing the Minimal Generator Family for Concept Lattices and Icebergs. In: B. Ganter and R. Godin (Eds.): ICFCA 2005, LNCS 3403, 2005: 192–207.

15. Kusnetsov SO. Stability as an estimate validity of hypotheses derived on the operational similarities basis. Scientific and technical information (Series 2), 1990; 12: 21-29. [In Russian]

16. Kusnetsov SO. On stability of a formal concept. Annals of Mathematics and Artificial Intelligence, 2007; 49(1-4): 101-115.

17. Zagoruyko NG. Applied methods of data and knowledge analysis [In Russian]. Novosibirsk: Sobolev Institute of Mathematics, SB RAS, 1999.

18. Freeman L. Cliques, Galois Lattices, and the Structure of Human Social Groups. Social Networks, 1996; 18: 173-187.

19. Kazanskiy NL, Popov SB. Machine Vision System for Singularity Detection in Monitoring the Long Process. Optical Memory and Neural Networks (Information Optics), 2010; 19(1): 23-30.

20. Ignatov DI, Kononychina ON. Formal concept lattices for data analysis in sociological interrogations. Integrated models and soft computation in Artificial Intelligence: Proc. of 5th Int. Conf. (Kolomna, Russia, 2009, May 20-30). Vol. 1. Moscow: "Fizmathlit" Publisher, 2009: 230-240. [In Russian]