Defuzzification of the initial context in Formal Concept Analysis

D E Samoilov^{1,2}, V A Semenova^{2,3} and S V Smirnov²

¹Samara National Research University, Moskovskoe Shosse, 34A, Samara, Russia, 443086
 ²Institute for Control of Complex Systems of RAS, Sadovaya Str., 61, Samara, Russia 443020
 ³Samara State Technical University, Molodogvardeyskaya Str., 244, Samara, Russia, 443100

e-mail: smirnov@iccs.ru

Abstract. The research field is the problem of extracting from the initial empirical material the formal concept lattice, which can serve as the basis of the formal ontology of the studied subject domain. The initial empirical material, i.e. the data of multidimensional observations and experiments, is characterized by incompleteness and inconsistency, conditioned by realities of empirical information accumulation. This leads to the fact that required for lattice building formal context can be previously presented only within the framework of some multivalued logic. It needs to be approximated in binary logic, since effective methods for derivation of formal concepts are developed only for unambiguous (binary) formal contexts. The exact solution of this problem, considering the properties existence constraints of objects in the studied subject domain, is difficult and in a certain sense is inadequate to expectations of subject exploring the subject domain. For defuzzification of the initial formal context heuristic was proposed, idea of which is to localize the approximation task of "soft" context within every group of dependent properties of each object of learning sample. The model reflecting such restrictions is formed as hierarchy of groups of dependent properties, which predetermines the recursive and multi-pass nature of the developed defuzzification algorithm.

1. Introduction

Standardly, protocols of observations and experiments are drawn up as tables "object-properties" (TOP) [1, 2]. Most often, TOP contains the results of measuring the values of the certain properties set of objects that are in the field of researcher view. Such TOPs are called multi-valued formal contexts (FCs) for data analysis tasks [3]. At the same time a very popular method of "object-properties" type data mining now is the formal concept analysis (FCA) [3-7], which deals only with binary (single-valued) FCs:

 (G^*, M, I)

where $G^* = \{g_i\}_{i=1,...,r}$, $r = |G^*| \ge 1$ – is the set of observed objects: $G^* \subseteq G$, G – all hypothetically conceivable set of objects of the studied knowledge domain (KD), $M = \{m_j\}_{j=1,...,s}$, $s = |M| \ge 1$ – set of measurable properties of the objects; $I \subseteq G^* \times M$ – the binary relation described by the incidence matrix "object-properties", each element of which is a truth estimate of basic semantic propositions (BSP) about the studied KD:

 b_{ij} = «oobject $g_i \in G^*$ has the property $m_i \in M$ »,

i.e. $I = (||b_{ij}||)_{i=1,...,r; j=1,...,s}, ||b_{ij}|| \in \{\text{Truth, False}\}.$

The main purpose of FCA is to extract clusters called formal concepts from tabular data. A partially ordered set of all formal concepts is called a lattice of concepts, which can be considered as a "skeleton" of the formal ontology of the studied KD [8-10].

A single-valued FC can be obtained by a certain granulation of the information of a multi-valued FC with help of the conceptual scales [11, 12]. Process of granulation depends on the objectives of the KD studying. This granularity generation tool is similar to the linguistic variables introduced by Zadeh [13]. Fuzzy conceptual scaling [14, 15] leads to FC of the form (1) with a fuzzy relation I. A slightly different approach to conceptual scaling that forms a fuzzy relation I was studied in [16].

In works [17-19], the genesis of a non-strict (in particular, fuzzy) FC (NFC) is considered from more general positions and is generally associated with incompleteness and inconsistency of the primary information about the studied KD. At the same time, it was established that fuzzy logic for describing "soft" correspondence of "object-properties" is less adequate than more sophisticated multi-valued logics. In [18, 19], in order to form the truth estimates of the BSP, it was proposed to use vector logic [20], namely, V^{TF} logic, which can be considered as the simplest generalization of Zadeh's fuzzy logic.

Constructive use of a NFC is based on its α -approximation, i.e. replacing in (1) fuzzy (or non-strict in the case of using vector logics) relation I with its α -section I^(α) [12, 14-19]. Unfortunately, in the general case a single-valued correspondence $I^{(\alpha)}$ in the problem of formal concept analysis turns out to be incorrect, because the standard α -section procedure does not take into account the existential dependencies between the measurable properties: incompatibility and conditionality of the objects properties of the studied KD [19, 21, 22]. A comprehensive model of a system of measurable properties (SMP) with restrictions on their existence was justified in [23, 24]. This work is devoted to the development of an algorithm for rational α -approximation of a NFC, which takes into account the complex organization of a SMP in FCA problems.

2. Truth evaluation of the propositions about the object properties

In V^{TF} logic the truth of BSP b_{ii} is estimated by the vector (Truth, False) (figure 1):

 $||b_{ij}|| = \langle b^+_{ij}, b^-_{ij} \rangle; b^+_{ij}, b^-_{ij} \in [0, 1]; (b^+_{ij} + b^-_{ij} \equiv 1) =$ **False**. Each object of the training sample $g_i \in G^*, i = 1, ..., r$, is characterized in a NFC by the set of truth vectors $\{\|b_{ij}\|\}_{j=1,...,s}$ (figure 2). V^{TF} -threshold of trust $\alpha = \langle \alpha^+, \alpha^- \rangle$ determines two regions in the existence space of truth vectors. Location of the vector $||b_{ij}||$ in one of these regions is interpreted as falsity, and for other region it is interpreted as truth of the corresponding BSP (figure 3). On the other hand, in the specified space for each vector $||b_{ij}||$ two regions are defined, location of the trust threshold α in one of them means falsity, and in the other - the truth of the BSP b_{ii} (figure 4).

3. Patterns of groups of conjugate properties

According to [23, 24], the SMP is a hierarchies set of groups of conjugate properties (GCP), which is formed as a product of subject hypothesizing about the structure of the conceptual description of the studied KD. Each GCP hierarchy is formed by replacing in a supergroup one or more measurable properties with a subgroup of such properties. The highest supergroup in the GCP nesting hierarchy is proto-property, the lower subgroups in the hierarchy are the "leafy" GCP, that are matched to measured properties.

In [23, 24] it was established that there are three and only three basic patterns of subgroups (subGCP) that replace one property in a supergroup:

- a group of two measurable properties, in which one property *conditiones* another;
- a group of n > 1 incompatible measurable properties;
- a group of n > 1 mutually conditioned properties. •



 $\mathbf{U} = \langle 0, 0 \rangle \qquad b^+_{ij} \qquad \mathbf{T} = \langle 1, 0 \rangle$ **Figure 1.** Representation of the truth vector $\|b_{ij}\| = \langle b^+_{ij}, b^-_{ij} \rangle \text{ and truth constants of } V^{TF} \text{ logic}$ «True», «False», «Uncertainty», «Contradiction», «Equivocal» (**T**, **F**, **U**, **C**, **E**).





Figure 2. Set of truth vectors $\{||b_{ij}||\}_{j=1,...,s}$, characterizing in a non-strict formal context BSP about object $g_i \in G^*$.



location of the vector $||b_{ij}||$ in this area means truth of corresponding BSP.

Figure 3. Determining of α -section; $\alpha = \langle \alpha^+, \alpha^- \rangle$ of a non-strict set using the logical order (or verisimilitude) for truth vectors in V^{TF} logic: $||b_{xy}|| > \alpha$, if $b^+_{xy} \ge \alpha^+$, $b^-_{xy} \le \alpha^-$.



Figure 4. Regions of choice of the trust threshold for the source data $\alpha = \langle \alpha^+, \alpha^- \rangle$, determining the approximate truth estimate in the scale {**True**, **False**} of the BSP, which vectorial truth estimation is $\langle b^+_{ij}, b^-_{ij} \rangle$.

The nature of properties conjugacy in such GCP can be clearly reflected in the existence space of truth vectors by connecting its points with appropriate arcs (figure 5a-c).

Similarly can be represented two complex patterns of subgroups, which replace several properties in a supergroup according to fixed rules (only in supergroups, where all properties are mutually conditioned):

- a group of properties with several conditionalities (figure 6a);
- a group of properties with several incompatibilities and conditionalities (figure 6b).

4. Rational *a*-section of a non-strict formal context

So, in order to apply effective FCA methods to extract formal concepts from a NFC, it is necessary to ensure its correct α -approximation in the presence of properties existence constraints (PEC).



Figure 5. Examples of combined diagrams of properties conjugacy and truth estimates of BSP corresponding to these properties for *basic* GCP patterns: (a) a group where one property conditiones another (C-group); (b) a group of incompatible properties (I-group); (c) a group of mutually conditioned properties (MUC-group).



Figure 6. Examples of combined diagrams of properties conjugacy and truth estimates of BSP corresponding to these properties for *mixed* GCP patterns (M-group): (a) a group that includes subGCP-pairs of properties with conditionalities (C-pairs); (b) a group including subGCP-pairs of properties with incompatibilities (I-pairs) and C-pairs.

4.1. Search of region of acceptable values of the trust threshold for the source data

Formally, the problem can be reduced to constructing a single predicate " α -section is correct" with a vectorial argument $\alpha = \langle \alpha^+, \alpha^- \rangle$, $\alpha^+, \alpha^- \in [0, 1]$, where condition for confirming the truth of each empirical BSP b_{ij} :

$$b^+_{ij} \ge \alpha^+ \wedge b^-_{ij} \le \alpha^-$$

(or, on the contrary, the necessary falsity of this BSP) should be combined with the implementation of all relevant PEC. And then find the region (possibly, it will be empty) of existence of trust thresholds α that deliver the value **True** to such predicate.

In the general case, to build such predicate and identify the specified region is very difficult; an example of constructing such predicate for the case when PEC is caused by incompatibility of some part of measurable properties can be found in [25]. However, even assuming the possibility of such a

decision, it is very *impractical* to bind the subject who studies the KD with a need to select a threshold only from *a limited region*. For example, with this approach the subject may not meet the intuitive expectations from softening or tightening the trust threshold for the data representing the KD [25].

4.2. *Heuristic approach*

Instead of the described search the following heuristic is proposed:

- the subject is free to choose a threshold (the choice of α -threshold is *arbitrary*);
- determined by a threshold α and, in general, the unacceptable composition of *each* object properties of the sought-for single-valued FC (in figure 2 this corresponds to the emphasizing of part of estimates $||b_{ij}||$ according to the rule illustrated by figure 3) should be consistently reduced due to the sequential cutting off of properties that violate PEC;
- the cut-off mechanism consists in local tightening of the trust threshold within each GCP in the SMP.

For unambiguous choice of "best" among trust thresholds - $\beta = \langle \beta^+, \beta^- \rangle$, - which tighten the threshold α chosen by the subject and provide the necessary reduction of the GCP composition of the object $g_i \in G^*$, following criterias are proposed (figure 7):

• *tightening vector length (in a rectangular metric)*

$$|\beta^+ - \alpha^+| + |\beta^- - \alpha^-| = (\beta^+ - \alpha^+) + (\alpha^- - \beta^-) \rightarrow \min;$$

• *truth region area* of the tightened trust threshold

$$(1 - \beta^+) \times \beta^- \rightarrow \max;$$

reliability [20] of the tightening vector ("in coordinates of the tightening")

$$t_{\beta}^{+} - t_{\beta}^{-} = (\alpha^{-} - \beta^{-}) - (\beta^{+} - \alpha^{+}) \rightarrow \max$$

The first and third criterias define two manifolds that are mutually perpendicular segments (figure 7). This fact guarantees an unambiguous choice of β at consistent application of criterias for the reduction of GCP members.



Figure 7. The vector tightening (bold arrow) which sets the shift of trust threshold $\alpha = \langle \alpha^+, \alpha^- \rangle$ by threshold $\beta = \langle \beta^+, \beta^- \rangle$; \square – the truth region of threshold α , \square – the truth region of threshold β .

The proposed heuristic method for obtaining the correct unambiguous approximation of a NFC is *effective*.

Indeed, the method is implemented for each object of the training sample *separately*, and at some step one of following conditions will arise:

- either the properties set of an object g_i begins to satisfy PEC (note that an empty set of properties satisfies such constraints, but then an object with such properties "set" must be qualified as unidentified by introducing a new property "unidentified object" into the soughtfor FC);
- or it will be stated that there is a *ineradicable contradiction* between the initial context and PEC: there are BSP about object g_i with the truth estimate $||b_{ij}|| = \langle 1, 0 \rangle$ (i.e. true in classical two-valued logic), but violating PEC.

4.3. Defuzzification algorithm of a non-strict formal context

The algorithmization complexity of proposed heuristics is due to the multi-hierarchical structural organization of the SMP [23, 24] - q.v. Section 3. At the same time, it is clear that the work begins with the input by a subject of the trust threshold for the source data $\alpha = \langle \alpha +, \alpha - \rangle$ that suits him and performing the standard α -section of the correspondence "object-properties" of a NFC. The algorithm core is the correction cycle of the obtained α -approximation of a NFC, performed for each object $g_i \in G^*$. Therefore, in the further description of "corrective steps" we will omit the mention of the object being processed, implying by default this uniqueness characteristic of the considered GCPs. But first we note the following:

- A GCP is excluded from the desired FC when all nested GCP (and all "leafy" sub-GCP which are matched to measured properties) are excluded from it. The local (for the GCP) trust threshold for the source data that implements such an *elimination* will be called the *absence*-, or GCP *a*-threshold (figure 8);
- A GCP will be included in the desired single-valued FC if it retains/preserves at least one GCP in it (and, therefore, at least one "leafy" sub-GCP associated with the measured property). The local (for the GCP) trust threshold for the source data that implements such an *inclusion* will be referred to as the GCP *presence-* or *p*-threshold (figure 9).

The NFC defuzzification will be completed when each GCP in the SMP will have its own value of *p*-threshold. For the object $g_i \in G^*$ in the desired single-valued FC, only those BSP will be recognized as true which are related to the measurable properties matched to the preserved "leafy" GCP.



Figure 8.Definition of GCP *a*-threshold ⟨a⁺, a⁻⟩:
□ – truth region of *a*-threshold;
○ – eliminated truth vectors of BSPs corresponding to the measured properties, which (in turn) are matched to the "leafy" subGCP of the excluded GCP.



Figure 9. Definition of GCP *p*-threshold ⟨p⁺, p⁻⟩:
□ – truth region of *p*-threshold;
● – eliminated and ⊗ – retained truth vectors of BSP corresponding to the measured properties, which (in turn) are matched to the "leafy" subGCP of the considered GCP.

4.3.1. Detection of GCP a-thresholds

At the first step of the correction cycle heuristically optimal a-threshold of each GCP is detected.

For the GCP, the recursive identification of all its "leafy" subGCPs mapped to the measured properties is implemented, and an actual set of truth vectors of the corresponding BSPs which are saved as a result of standard NFC α -section is generated.

Of course, the actual set of GCP truth vectors may turn out to be empty (figure 10), and, therefore, the considered GCP is excluded from the desired FC as a result of the standard α -section of a NFC. In this case, the identification of the optimal GCP a-threshold does not make sense, and some "disabling" of the GCP *p*-threshold may be a sign of this GCP cut-off. Assigning to *p*-threshold abscissa a negative value may serve as example of such disabling.

The heuristically optimal GCP *a*-threshold can be detected only with a non-empty actual set of GCP truth vectors, and, as it is not difficult to understand, the members of the southeast Pareto-front of this set will serve as reference vectors for determining the optimal GCP *a*-threshold (figure 11). GCP *a*-threshold itself should be searched (for example, by brute force) at the intersection of lines which are parallel to the coordinate axes and passing through the reference vectors (see rhombs in figure 11):

- intersections of lines $x = b_X^+$ and $y = b_Y^-$, respectively, of the first and second members of each pair of neighboring truth vectors $\langle b_X^+, b_X^- \rangle \bowtie \langle b_Y^+, b_Y^- \rangle$ of the Pareto-front;
- the intersection of the $x = b_N^+$ line of the "northern" truth vector $\langle b_N^+, \bar{b}_N^- \rangle$ of the Pareto-front with the $y = \alpha^-$ line;
- the intersection of the $y = b_s^-$ line of the most "southern" the truth vector $\langle b_s^+, b_s^- \rangle$ of the Pareto-front with the $x = \alpha^+$ line.

An intersection that is "best" in terms of the proposed criterias is accepted as the heuristically optimal GCP *a*-threshold. In the further analysis of the data, it is taken into account that indeed optimal GCP *a*-threshold is more plausible, but it infinitely small differs from the heuristically optimal one.

4.3.2. Detection of GCP p-thresholds

At the second (and final) step of the correction cycle, *a heuristically optimal p*-threshold of *each* GCP is detected. The main work is performed by:

- a recursive procedure of local GCP correction (with the detection of its heuristically optimal *p*-threshold). One of the procedure input parameters is the "current *p*-threshold";
- a recursive procedure of GCP exclusion from the desired FC (with the disabling of its *p*-threshold).





Figure 10. Elimination of GCP from the desired FC as a result of the standard α -section of the source context (the designation of the region and the truth vectors is the same as in figures 7 and 8).



The first procedure is sequentially launched for the "root" GCP with the current threshold equal to α . Recursively, possibly with reiteration of passes, it performs correction of the "root" GCP and all its subGCP. The second procedure plays an auxiliary role in execution of the first one, but exactly at an attempt to exclude the GCP ineradicable contradiction of the initial context and PEC can be found (q.v. Subsection 4.2).

Let's explain these actions.

In relation to any GCP, the trust threshold for the data $\beta = \langle \beta^+, \beta^- \rangle$ either excludes or retains it in the desired FC.

In the first case, GCP *p*-threshold is disabled (i.e., the GCP exclusion from the desired FC is fixed) and, if the GCP is not "leafy", *p*-thresholds of all embedded subGCPs of the considered GCP are disabled.

In the second case, if the given GCP is "leafy", then the *p*-threshold is equal to β . Otherwise, the execution of local PEC is checked at threshold β :

- first, all embedded sub-GCPs of the considered GCP are corrected recursively (with detection of *p*-thresholds);
- secondly, *according to the rules that depend on the type* of a given GCP, its *p*-threshold is revealed.

Here are these rules:

- GCP MUC-group (see figure 5c). If at least one subGCP has a *p*-threshold disabled, then *p*-threshold of the MUC-group and *p*-thresholds of all its subGCP are disabled. Otherwise, the *p*-threshold of the MUC-group is set to β;
- GCP C-group or C-pair (see figure 5a or figure 6). If the conditioned subGCP is excluded, then the *p*-thresholds of the C-group/C-pair and the conditioning subGCP are disabled. Otherwise, the *p*-threshold of the C-group/C-pair is set to β;
- GCP I-group or I-pair (see figure 5 and figure 6b). All non-excluded subGCP are incompatibility *violators*. From violators we select subGCP with a "best" (from the position of the proposed criterias) *a*-threshold, which is assigned as new β. Correction of the considered I-group/I-pair is repeated. Reiterations continue as long as two or more non-excluded subGCPs retain. As a result, the *p*-threshold of the I-group/I-pair will be either disabled or equated to the current β.
- GCP M-group (see figure 6). Members which violates conditionality and incompatibility (i.e. certain subsubGCP) are selected from non-excluded subGCP-pairs. Then subsubGCP with the "best" *a*-threshold is selected from them and this a-threshold is assigned as new β. Correction of the considered M-group is repeated. Reiterations continue as long as there are violators. As a result, the *p*-threshold of the M-group will be either disabled or equated to the current β.

5. Conclusion

The proposed heuristic method and the developed algorithm of defuzzification of the initial context allows one to take into account the complex organization of the SMP of the studied KD in problems of the FCA. The obtained result "closes" the main section of the hypothetical-deductive theory of ontological data analysis developed by authors [17-19, 23-25].

It is important to study the effectiveness of the developed algorithm depending on the parameters characterizing the SMP of the studied KD and the initial non-strict context of FCA problem.

The presented method and algorithm are focused on the unambiguous approximation of non-strict formal contexts and the derivation of "crisp" concepts. A promising task is the development of an appropriate method and algorithm, the end result of which is a system of "fuzzy" concepts.

6. References

- [1] Barsegyan A A, Kupriyanov M S, Holod I I, Tess M D and Elizarov S I 2009 *Data and Process Analysis* (St. Petersburg: BHV-Petersburg)
- [2] Zagoruyko N G 2013 *Cognitive data analysis* (Novosibirsk: Sobolev Institute of Mathematics, SB RAS)
- [3] Ganter B and Wille R 1999 Formal Concept Analysis. Mathematical foundations (Springer-Verlag Berlin-Heidelberg)
- [4] Formal Concept Analysis Homepage URL: http://www.upriss.org.uk/fca/fca.html (15.05.2019)
- [5] Ganter B and Obiedkov S 2016 *Conceptual Exploration* (Springer-Verlag Berlin Heidelberg)
- [6] Ignatov D I 2015 Introduction to Formal Concept Analysis and Its Applications in Information Retrieval and Related Filds Information Retrieval. Revised Selected Papers 8th Russian Summer School (Nizhniy Novgorod, Russia, Springer International Publishing) 42-141
- [7] Carpineto C and Romano G 2004 *Concept Data Analysis: Theory and Applications* (Wiley)
- [8] Obitko M, Snasel V and Smid J 2004 Proc. of the CLA 2004 International Workshop on Concept Lattices and their Applications (Ostrava, Czech Republic) 111-119
- [9] Smirnov S V 2001 Ontological analysis of modeling domain *Bulletin of the Samara Scientific Center of RAS* **3** 62
- [10] Godin R, Mili H, Mineau G W, Missaoui R, Arfi A and Chau T T 1998 Ontology Design with Formal Concept Analysis *Theory and Application of Object Systems (TAPOS)* **4(2)** 117
- [11] Ganter B and Wille R 1989 Conceptual scaling *Applications of Combinatorics and Graph Theory to the Biological and Social Sciences* (New York Springer-Verlag) 139
- [12] Wolff K E 1998 Conceptual Interpretation of Fuzzy Theory 6th European Congress on Intelligent Techniques and Soft Computing (Germany, Aachen) **1** 555-562

- [13] Zadeh L A 1975 The concept of a linguistic variable and its application to approximate reasoning *Information Science* **8** 199 (Part I), **8** 301 (Part II), **9** 43 (Part III)
- [14] Belohlavek R and Konecny J 2007 Scaling, Granulation, and Fuzzy Attributes in Formal Concept Analysis *The IEEE International Conference on Fuzzy Systems* (London) 918
- [15] Yang K M, Kim E H, Hwang S H and Choi S H 2008 Fuzzy Concept Mining based on Formal Concept Analysis Int. J. of Computers 2 279
- [16] Pollandt S 1996 Fuzzy-Begriffe: Formale Begriffsanalyse unscharfer Daten (Springer-Verlag Berlin Heidelberg)
- [17] Smirnov S V 2007 Fuzzy formal contexts in building of ontologies based on the formal concepts analysis: genesis and use *Proc. All-Russian Conf. "Knowledge–Ontologies–Theories"* (Novosibirsk, Sobolev Institute of Mathematics, SB of RAS) 2 17-25
- [18] Smirnov S V 2014 Multi-valued and fuzzy logic in ontological data analysis Proc. 3rd Int. scientific conf. "Information technology and systems" (Publishing House of Chelyabinsk State University) 90-91
- [19] Semenova V A and Smirnov S V 2016 Intelligent analysis of incomplete data to building formal ontologies *CEUR Workshop Proceedings* **1638** 796-805
- [20] Arshinskii L V 2007 Substantial and formal deductions in logics with vector semantics *Automation and Remote Control* **68(1)** 139
- [21] Pronina V A and Shipilina L B 2009 Using relationships between attributes to build a domain ontology *Control sciences* **1** 27
- [22] Lammari N and Metais E 2004 Building and maintaining ontologies: a set of algorithms Data & Knowledge Engineering 48(2) 155
- [23] Samoylov D E, Semenova V A and Smirnov S V 2018 Multilevel recursive model of properties existence constraints in machine learning *Journal of Physics: Conf. Series* **1096** 012096
- [24] Samoylov D E, Semenova V A and Smirnov S V 2018 The structure of measured properties system for objects of multidimensional observation and experiments *Bulletin of the Samara State Technical University Series "Technical Sciences"* **3(59)** 56
- [25] Ofitserov V P, Smirnov V S and Smirnov S V 2014 Method of the alpha-section of non-strict formal contexts in the formal concepts analysis *Proc. of the 16th Int. Conf. Problems of Control and Modeling in Complex Systems* (Samara: Samara Scientific Center of RAS) 228