УДК 519.7

On Decomposition of a Binary Context Without Losing Formal Concepts

Choduraa M. Mongush*
Institute of Mathematics and Computer Science
Siberian Federal University
Svobodny, 79, Krasnoyarsk, 660041
Tuvan State University
Lenina, 36, Kyzyl, Tuva, 667000
Russia

Valentina V. Bykova†
Institute of Mathematics and Computer Science
Siberian Federal University
Svobodny, 79, Krasnoyarsk, 660041
Russia

Received 11.11.2018, received in revised form 11.02.2019, accepted 10.03.2019

The #P-complete problem of finding all the formal concepts of a given context and the decomposition method for its solving are investigated. As parts of the decomposition is proposed to use fragments of the initial context, called boxes. Such decomposition allows to decompose the given context without losing formal concepts and thereby to reduce the execution time of the algorithms for solving considered task. The number of boxes, obtained at each iteration of the decomposition, is determined based on studies of the boxes structure and the rules for stopping of the decomposition process are established.

Keywords: formal concept analysis, decomposition of the binary context.
DOI: 10.17516/1997-1397-2019-12-3-323-330.

Introduction

Initial data is often represented as an object-attribute table in data mining tasks, in which each column corresponds to a certain attribute, and each row defines a characteristic description of an separate object [1]. For example, within the formal concept analysis (FCA) an object-attribute table is modelled by binary context, reflecting the presence or absence of attributes specific to the studied set of object. Formal concept analysis as a direction in mathematics arose in the early 80s of the last century with the advent of the works of R. Wille and B. Ganter [2]. In FCA the generally accepted term "concept" is formalized and defined by a pair of the sets (extension, intention), called a formal concept.

The various data mining tasks such as the classification, conceptual modeling a subject domain, identifying dependencies among data can be solved with the help of FCA [3–6]. The solution of these problems is reduced to compute all the formal concepts of the considered subject domain. However, the process of generation all formal concepts has a high computational

*mongushchod91@yandex.ru
†bykvalen@mail.ru

complexity. Since the number of formal concept can be exponential from size of the initial context [2]. The task of finding all formal concepts has been studied in detail in a variety of published papers [3–5]. It is also proved that it belongs to the class of $\#P$-complete problems [2, 4]. The problem of high computational complexity of finding all formal concepts significantly complicates the practical application of FCA in data mining.

One of the approaches of solution this problem was described in the paper [6], where was proposed the decomposition method which allows to decompose the initial context into parts (boxes) without losing formal concepts and thereby to reduce the time of finding all the formal concepts of this context. In this paper, the boxes structure is theoretically investigated, the rules for organization of the iterative process of forming a set of boxes are presented, the reception for reducing the number of boxes at each separate iteration of the decomposition is proposed.

1. The main provisions and notations

Let's consider the basic definitions and typical notations of FCA [2, 3].

Let two non-empty finite sets are defined: a set of objects G (*Gegenstände*) and a set of attributes or properties M (*Merkmale*). Let there be also given non-empty incidence relation $I \subseteq G \times M$. This relation contains information about the satisfiability of properties from M on objects from G, i.e. $(g, m) \in I$ means that the object g has property m and conversely, the attribute or property m is inherent to object g. A triple $K = (G, M, I)$ is called a formal context.

Further let us assume that the sets G and M are linearly ordered (for example, lexicographical). Then the context $K = (G, M, I)$ is uniquely given by the 0,1-matrix $T = (t_{ij})$: $t_{ij} = 0$ at $(g_i, m_j) \notin I$ and $t_{ij} = 1$ at $(g_i, m_j) \in I$ $(i = 1, 2, \ldots, |G|;\ j = 1, 2, \ldots, |M|)$.

Let us choose in $K = (G, M, I)$ two arbitrary elements $g \in G$, $m \in M$ and define for them the mappings $(\cdot)'$ as follows:

$$g' = \{m \in M \mid (g, m) \in I\}, m' = \{g \in G \mid (g, m) \in I\}. \tag{1}$$

According to this definition, the set g' is a set of attributes inherent to the object $g \in G$, and m' is a set of objects that have attribute $m \in M$. The mappings $(\cdot)'$ are easily generalized to the sets $A \subseteq G$ and $B \subseteq M$: $A' = \cap_{g \in A}\, g'$, $B' = \cap_{m \in B}\, m'$.

For each formal context $K = (G, M, I)$ and any subsets $B_1, B_2 \subseteq M$ the properties are correct:

- *antimonotony*: if $B_1 \subseteq B_2$, then $(B_2)' \subseteq (B_1)'$;
- *extensiveness*: $B_1 \subseteq (B_1)''$, where $(B_1)'' = ((B_1)')' \subseteq M$.

Similar properties are fair for subsets of a set G. It is known that for the mappings $(\cdot)'$ the equalities are also true:

$$\big((A')'\big)' = (A'')' = A', ((B')')' = (B'')' = B'. \tag{2}$$

The double application of the mapping $(\cdot)'$ defines a closure operator $(\cdot)''$ to 2^M in the algebraic sense. This operator following properties are fair:

- *reflexive*: for any $B \subseteq M$ always $B \subseteq B''$;
- *monotony*: if $B_1 \subseteq B_2 \subseteq M$, then $(B_1)'' \subseteq (B_2)'' \subseteq M$;
- *idempotency*: for any $B \subseteq M$ always $(B'')'' = B''$.

A set $(B)''$ can be interpreted as a set of attributes that always appear in objects of the context of $K = (G, M, I)$, together with attributes from B and this set is the greatest by inclusion within this context. If $B = B''$, then B is called the closed set relatively of the operator $(\cdot)''$.

A pair of sets (A, B), $A \subseteq G$, $B \subseteq M$, such that $A' = B$ and $B' = A$, is called a formal concept of the formal context $K = (G, M, I)$ with an extent A and an intent B. Further, in some cases the definition "formal" will shall omit before the words "context" or "concept".

From (2) and the definition of the operator $(\cdot)''$ follows the validity of the statement: a pair of sets (A, B) is the formal concept if and only if $A = A''$ and $B = B''$. Obvious also that every formal concept is unique in a given context, i.e. it differs from other formal concepts by its extent and/or intent. If a context $K = (G, M, I)$ is represented by the 0,1-matrix T, then a formal concept (A, B) at $A \neq \varnothing$ and $B \neq \varnothing$ is corresponded the maximal full sub-matrix of the matrix T. The rows of this sub-matrix are corresponded to the elements from A, and the columns are corresponded to the elements from B. Here, the maximal full sub-matrix is a sub-matrix in which all elements are equal to 1 and which are not contained in other full sub-matrices.

We denote by FC a set of all formal concepts of the context $K = (G, M, I)$. Let $(A_1, B_1), (A_2, B_2) \in FC$. A set FC is partial ordered set with relation $(A_1, B_1) \sqsubseteq (A_2, B_2)$ if and only if $A_1 \subseteq A_2$. Note that the latter is equivalent to the condition $B_2 \subseteq B_1$. Each formal concept $(A, B) \in FC$ defines a set of homogeneous objects A of with its specific set of attributes B for studied subject domain. If in $K = (G, M, I)$ there is no attributes shared by all objects from G, then a set of FC will contain the formal concept $(G, \varnothing) \in FC$. If in $K = (G, M, I)$ there is no objects which have all the attributes from M, then $(\varnothing, M) \in FC$. If both cases are satisfied simultaneously, then $(G, \varnothing) \in FC$ and $(\varnothing, M) \in FC$. The formal concepts $(G, \varnothing), (\varnothing, M)$ is often called trivial.

2. On problem of complexity computing all formal concepts

In the task of computing all formal concepts it is required to find a set of FC for a given context $K = (G, M, I)$. As noted in the introduction, this task is $\#P$-complete. Currently for solution this problem are known the algorithms NextClosure, Close-by-One, Norris et al. The execution time of these algorithms in the worst case is $O(|FC| \cdot |G|^2 \cdot |M|)$ [3–5]. Nowadays the studies are underway to improve the performance of algorithms of finding all formal concepts due to parallel computing and of using the decomposition approach [3, 6]. The main goal of these studies is to make the methods of FCA more accessible for analyzing big data sets.

A decomposition approach for solving the task of computing all the formal concepts of a given context is its reduction to a finite series of subtasks. Each of these subtasks is a reduced copy of the original task, which is solved on a certain part of the given context. The decomposition process is aimed at consistently reducing the size of parts of the context. As a result, a finite set of different parts is formed (in the general case of different sizes and having a non-empty intersection).

The decomposition process is implemented iteratively, since recursion in such cases is more time consuming [7]. For the organization of the decomposition process is required to determine: the rule decomposition of the context (what are parts and how to distinguish them); the estimate the number of parts obtained at each iteration of the decomposition; the rule stopping of the decomposition process. In addition, for any decomposition method of solving a problem, the rules of renewal the desired solution are required, based on the solutions obtained for the subtasks. The polynomiality on time of the separation procedures of the input data of the solving problem

into parts is the main requirement which when performing this requirement the decomposition effect is achieved. We further describe the essence of the decomposition method of the formal context proposed in [6].

3. The decomposition process of the formal context

Let $K = (G, M, I)$ be a context, FC be a set of all its formal concepts and T be a 0,1-matrix corresponding to it.

The context $K_1 = (G_1, M_1, I_1)$ is called a part of $K = (G, M, I)$, if $G_1 \subseteq G$, $M_1 \subseteq M$ and for any $x \in G_1$, $y \in M_1$ the relation $(x, y) \in I_1$ is true if and only if $(x, y) \in I$. The context $K_1 = (G_1, M_1, I_1)$ corresponds to a submatrix of the matrix T, in which the rows corresponding to the objects from $G \setminus G_1$ and the columns corresponding to the attributes from $M \setminus M_1$ are deleted. Any nontrivial formal concept from FC can be considered as part of the context $K = (G, M, I)$. The parts $K_1 = (G_1, M_1, I_1)$ and $K_2 = (G_2, M_2, I_2)$ are called different if $G_1 \neq G_2$ and/or $M_1 \neq M_2$.

It is required to decompose the context $K = (G, M, I)$ a finite set of different parts so that the conditions are fulfilled:

- each part contains at least one formal concept from FC;
- neither one formal concept from FC is lost, and new formal concepts do not arise, i.e. concepts that are not in FC.

A decomposition satisfying conditions 1 and 2 is called "safety" relatively of the formal concepts. If the 0,1-matrix T is full, then the resulting set will consist of only one part representing the context $K = (G, M, I)$, and this part will contain only one formal concept (G, M). Obviously that the greatest number of different parts into which the context can be "safety" decomposed is equal $|FC|$, i.e. the number of formal concepts of the context $K = (G, M, I)$. Since there are contexts for which the number of formal concepts exponentially depends on $|G|$ and $|M|$, then the number of parts obtained at each iteration of the decomposition is advisable to estimate and to determine the stopping rule for the implementation of the all decomposition process in polynomial time. At first we give the main provisions of the work [6].

Let $g \in G$ and $m \in M$ is arbitrary elements of the context $K = (G, M, I)$. Pairs of sets (g'', g') and (m', m'') form formal concepts the first of which is called the object concept, and the second is the attribute formal concept of the context $K = (G, M, I)$.

Let us denoted by $O = \{(g'', g') \mid \forall g \in G\} \subseteq FC$ the set of all object concepts and by $S = \{(m', m'') \mid \forall m \in M\} \subseteq FC$ the set of all attribute concepts of the context $K = (G, M, I)$. A pair of the formal concepts $(g'', g') \in O$ and $(m', m'') \in S$ defines the box (m', g', J) as part of the context $K = (G, M, I)$, if

$$(g'', g') \sqsubseteq (m', m''), \tag{3}$$

which is equivalent to $g'' \subseteq m'$ (or $m'' \subseteq g'$). About same box say that it formed by the elements $g \in G$ and $m \in M$. Further, instead of (m', g', J) we will briefly write (m', g').

Let us say that a formal concept $(A, B) \in FC$ is embedded into the box (m', g') of context $K = (G, M, I)$, symbolically $(A, B) \preceq (m', g')$, if $A \subseteq m'$, $B \subseteq g'$. Any box (m', g') is not empty, since according to (3) it always contains concepts $(g'', g') \in O$, $(m', m'') \in S$.

The correspondence between the boxes and the formal concepts of context $K = (G, M, I)$ establishes the following theorem which proved in [6].

Theorem 1. *For each context $K = (G, M, I)$, a set FC of all its formal concepts and any pair of sets (A, B), $\varnothing \neq A \subseteq G, \varnothing \neq B \subseteq M$, are fair the following statements:*

1. *if $(A, B) \in FC$, then always in context $K = (G, M, I)$ there is a box (m', g'), $g \in G$ and $m \in M$, at that perhaps not the only one in which this formal concept is embedded;*
2. *if (A, B) is the formal concept of certain box (m', g') of the context $K = (G, M, I)$, then it belongs FC.*

According to Theorem 1, the decomposition of the context $K = (G, M, I)$ on boxes is "safety" for any formal concept from FC. An exception is the case when FC contains at least one of the trivial formal concepts $(G, \varnothing)$, $(\varnothing, M)$. Since relations are always true

$$(\varnothing, M) \sqsubseteq (G, \varnothing), (\varnothing, M) \sqsubseteq (G, G'), (M', M) \sqsubseteq (G, \varnothing),$$

then the context $K = (G, M, I)$ can be considered as a box (G, M). Therefore, and even in these exceptional cases each box contains at least one formal concept from FC, herewith neither one formal concept from FC is not lost.

From Theorem 1 implies an important practical consequence: the desired set FC can be reconstructed by combining the sets of formal concepts revealed in boxes of the formal context $K = (G, M, I)$.

Note that the decomposition process of the context on the boxes can be organized iteratively. Since each box, identified at the first iteration, is considered as the initial context and again undergoes to decomposition. We investigate the boxes structure to estimate the number of boxes obtained at each iteration of the decomposition and to determine the rule for stopping the decomposition.

4. A study of boxes structure

Proposition 2. *For each context $K = (G, M, I)$ and any $(g'', g') \in O$, $(m', m'') \in S$ the order relation $(g'', g') \sqsubseteq (m', m'')$ is satisfied if and only if $(g, m) \in I$.*

Proof. Let the order relation $(g'', g') \sqsubseteq (m', m'')$ is true. Then $g'' \subseteq m'$, $m'' \subseteq g'$. According to the reflexivity of the closure operator $(\cdot)''$, we have $\{g\} \subseteq g'' \subseteq m'$, $\{m\} \subseteq m'' \subseteq g'$. From (1) follows $(g, m) \in I$. Let us prove the opposite. Let $(g, m) \in I$. It means that $\{g\} \subseteq m'$, $\{m\} \subseteq g'$. By virtue the monotony of the closure operator $(\cdot)''$, the inclusions $g'' \subseteq (m')''$, $m'' \subseteq (g')''$ are fair. Hence, by the reflexivity of the closure operator $(\cdot)''$ and equalities (2), we have

$$\{g\} \subseteq g'' \subseteq m', \{m\} \subseteq m'' \subseteq g'.$$

Therefore, $(g'', g') \sqsubseteq (m', m'')$. Proposition 2 is proved. □

From Proposition 2 follows that the number of boxes generated by the various elements of the context $K = (G, M, I)$ is equal to the weight of the 0,1-matrix T, i.e. the quantity of $\|T\|$ is the number of unit elements of this matrix. Obviously that

$$1 \leqslant \|T\| \leqslant |G| \cdot |M|.$$

Proposition 3. *Any nontrivial formal concept (A, B) of the context $K = (G, M, I)$, which is embedded into the box (m', g'), formed by the elements $g \in G$ and $m \in M$, necessarily contains these elements and their closures, i.e. if $(A, B) \preceq (m', g')$, then always*

1) $g \in A$ *and* $m \in B$*;*

2) $g'' \subseteq A$ *and* $m'' \subseteq B$*.*

Proof. If $(A, B) \preceq (m', g')$ is true, then $A \subseteq m'$, $B \subseteq g'$. By virtue the antimonotony of the mapping $(\cdot)'$ we have $m'' \subseteq A'$, $g'' \subseteq B'$. For each formal concept (A, B) by definition is true $A = B'$, $B = A'$. Then

$$m'' \subseteq B, g'' \subseteq A. \tag{4}$$

By virtue the reflexivity of the closure operator $(\cdot)''$

$$\{m\} \subseteq B, \{g\} \subseteq A. \tag{5}$$

From relationships (4), (5) directly follows the justice of both statements of Proposition 3. □

According to Proposition 3, a pair of (g'', m'') can be considered as typical representatives not only of the box (m', g'), but also of all formal concepts of the context $K = (G, M, I)$ which is embedded into this box. This is correct, since the sub-matrix corresponding to the box (m', g') has the unit elements in all rows from g'' and all columns from m''. Meanwhile, there may be zero elements in it.

We introduce the notion of the box density. Let $|m'| \cdot |g'|$ be the size of box (m', g') and $\|(m', g')\|$ be the number of unit elements in this box. The box density (m', g') is called the quantity

$$\sigma\,(m', g') = \frac{\|(m', g')\|}{|m'| \cdot |g'|}.$$

For box density is true the natural boundaries $0 < \sigma\,(m', g') \leqslant 1$.

Proposition 4. *If the box* (m', g')*, formed by the elements* $g \in G$ *and* $m \in M$*, has the density* $\sigma\,(m', g') = 1$*, then* $g'' = m'$*,* $m'' = g'$*.*

Proof. Hence $\sigma\,(m', g') = 1$, then by definition of the box density is true $|m'| \cdot |g'| = \|(m', g')\|$. This means that for any object $g \in m'$ and for any attribute $m \in g'$ is true $(g, m) \in I$. Hence, $g' = m''$, $m' = g''$. □

Proposition 5. *Any box* (m', g') *with density* $\sigma\,(m', g') = 1$ *contains exactly one nontrivial formal concept* (A, B) *of the context* $K = (G, M, I)$ *that coinciding with it, i.e. the equalities* $A = m'$ *and* $B = g'$ *are true.*

Proof. Let $(A, B) \preceq (m', g')$ is true, then $A \subseteq m'$, $B \subseteq g'$. From Proposition 4 follows that $A \subseteq m' = g''$ and $B \subseteq g' = m''$, which means $A \subseteq g''$ and $B \subseteq m''$. Meanwhile, according to Proposition 3, reverse inclusions $g'' \subseteq A$ and $m'' \subseteq B$ are true. Therefore $A = m'$, $B = g'$. □

From Propositions 5 follows that the box (m', g') with density 1 degenerates into a nontrivial formal concept and is not subject to further decomposition.

Note that the time to build one box is $O\,(|G| \cdot |M|)$. According to Proposition 2, the number of boxes arising at each separate iteration of the decomposition process is comparable to $O\,(|G| \cdot |M|)$. If the number of iterations is limited to a certain constant, then in general the decomposition process of the initial context into boxes can be carried out in polynomial time.

In practice, the number of boxes arising at each separate iteration of the decomposition process can in some cases be reduced, since nested and coinciding boxes are possible among them.

Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_{\|T\|}\}$ be a set of boxes of the formal context $K = (G, M, I)$, where $\omega_i = (m_i', g_i')$, $i = 1, 2, \ldots, \|T\|$. Let us consider two boxes $\omega_1 = (m_1', g_1')$, $\omega_2 = (m_2', g_2') \in \Omega$. Let's say that the box $\omega_1 = (m_1', g_1')$ is embedded into the box $\omega_2 = (m_2', g_2')$, symbolically $\omega_1 \preceq \omega_2$, if set-theoretic inclusions are true:

$$m_1' \subseteq m_2' \text{ and } g_1' \subseteq g_2'.$$

If $m_1' = m_2'$ and $g_1' = g_2'$, then boxes ω_1 and ω_2 are called coinciding. The boxes ω_1 and ω_2 are comparable, if $\omega_1 \preceq \omega_2$ or $\omega_2 \preceq \omega_1$, otherwise they are incomparable. Thus, a set Ω is partially ordered relative of the order relation introduced above. In view of Theorem 1 following corollary is fair.

Corollary 6. *For any $\omega_1, \omega_2 \in \Omega$, such that $\omega_1 \preceq \omega_2$, following is true: all formal concepts of the box ω_1 are the formal concepts of the box ω_2 and consequently the formal concepts of the context $K = (G, M, I)$.*

It is known that in a partially ordered set always can find mutually disjoint chains. A non-empty subset $\{\omega_{i1}, \omega_{i2}, \ldots, \omega_{il}\}$ of a set Ω is a chain, if all elements of this subset are pairwise comparable and linearly ordered: $\omega_{i1} \preceq \omega_{i2} \preceq \ldots \preceq \omega_{il}$. The element ω_{il} is called the maximal element, and the value l is the length of this chain. A chain is maximal, if its union with any element, not belonging to it, is not a chain. Two chains are called mutually disjoint, if they do not contain common elements. The number of maximal mutually disjoint chains and the length of a longest such chain is determined by Dilworth's theorem [8]. The algorithm for constructing mutually disjoint chains of a partially ordered set based on the calculation of the maximum matching of a bipartite graph exists. In the paper [8] it was proved that the execution time of this algorithm is polynomial relatively of the power of the initial partially ordered set. It is also proved that the constructed chains are maximal and mutually disjoint.

According to Corollary 6, the maximal element of any chain preserves all the formal concepts of the other elements of this chain. These elements of a chain can be removed and by that the number of boxes obtained at each separate iteration of the decomposition is reduced. However, there are cases where the specified reception does not give effect. For example, in the case when all elements of the set Ω are pairwise incomparable or when a set Ω is linearly ordered.

Conclusion

The decomposition method of presented in this paper allows to improve the performance of algorithms for solving the problem of finding all formal concepts, and to apply them for subject areas described by contexts with big dimension. Other decomposition methods of the context on the part are possible, but they should always be "safety" relatively of the formal concepts.

References

[1] A.A.Barseghyan, M.S.Kupriyanov, V.V.Stepanenko, I.I.Kholod, Data Analysis Technology: Data Mining, Visual Mining, Text Mining, OLAP, Piter, St.Peterburg, 2008 (in Russian).

[2] B.Ganter, R.Wille, Formal Concept Analyses: mathematical foundations, Springer Science and Business Media, 2012.

[3] B.Ganter, S.A.Obiedkov, Conceptual Exploration, Berlin, Heidelberg, Springer, 2016.

[4] S.O.Kuznetsov, S.A.Obiedkov, Comparing Performance of Algorithms for Generating Concept Lattices, *Journal of Experimental and Theoretical Artificial Intelligence*, **14**(2002), no. 2–3, 189–216.

[5] A.A.Simon, "Best-of-Breed" approach for designing a fast algorithm for computing fixpoints of Galois Connections, *Information Sciences*, **265**(2015), 633–649.

[6] V.V.Bykova, Ch.M.Mongush, On Algebraic Approach of R. Wille and B. Ganter in the Investigation of Texts, *Journal of Siberian Federal University*, **10**(2017), no. 3, 372–384.

[7] V.V.Bykova, Mathematical methods for analyzing recursive algorithms, *Journal of Siberian Federal University. Mathematics and Physics*, **3**(2008), no. 1, 236–246 (in Russian).

[8] E.Harzheim, Ordered Sets, New York, Springer, 2015.

О декомпозиции бинарного контекста без потери формальных понятий

Чодураа М. Монгуш
Институт математики и фундаментальной информатики
Сибирский федеральный университет
Свободный, 79, Красноярск, 660041
Тувинский государственный университет
Ленина, 36, Кызыл, Тыва, 667000
Россия

Валентина В. Быкова
Институт математики и фундаментальной информатики
Сибирский федеральный университет
Свободный, 79, Красноярск, 660041
Россия

Исследованы #P-полная задача нахождения всех формальных понятий заданного контекста и декомпозиционный метод ее решения. В качестве частей разложения предложено использовать фрагменты исходного контекста, названные боксами. Такая декомпозиция позволяет разлагать заданный контекст без потери формальных понятий и тем самым снижать время выполнения алгоритмов решения рассматриваемой задачи. На основе исследования структуры боксов определено число боксов, получаемых на каждой итерации разложения, и установлены правила остановки процесса разложения.

Ключевые слова: анализ формальных понятий, декомпозиция бинарного контекста.