Full polynomial probabilistic FCA-based knowledge extraction

Dmitry Vinogradov[1,*,†]

[1]Dorodnicyn Computing Center, Federal Research Center "Computer Science and Control", Russian Academy of Sciences, 40 Vavilova St., Moscow, 119333, Russian Federation

Abstract

The article demonstrates computational efficiency of the probabilistic approach to knowledge extraction using the FCA. In addition to the result previously proved by the author on sufficiency of a polynomial number of hypotheses (concepts) about the causes of the target property under study, this paper will give a polynomial upper bound on the average running time of the algorithm for generating one concept. The proven result concerns a family of algorithms based on coupling Markov chains for arbitrary formal contexts formed from the positive part of training sets. To get a good estimate for the length of trajectory (before entering to some ergodic state) of such a chain, we had to enrich the representation of the training sample by adding negation for every original binary attribute.

Keywords

formal concept, coupling Markov chain, mean length of trajectory, computational complexity

1. Introduction

The extraction of knowledge using a binary similarity operation began in the early 1980s in the works of Prof. V.K. Finn, who proposed the JSM-method of automatic generation of hypotheses [1, 2].

This approach was named after the English philosopher, economist and logician John Stuart Mill, whose ideas on Inductive Logic [3] served as the starting point of the JSM-method. The key component of this approach is a binary similarity operation. In the beginning, this operation was considered in isolation: most often as the intersection of sets of binary attributes describing training examples. In this case, it was a way of finding a set of common attributes. Initially, domains of objects (training and test examples) were Boolean algebras.

Then S.O. Kuznetsov proposed [4] to apply Formal Concept Analysis [5] to JSM-paradigm. This discovery led to extension of domains of application by those, that can be described by general lattices, and to invention of more efficient algorithms [6].

However, the JSM-method has a number of significant limitations that do not allow it to cope with training samples of moderate size. One of them is exponential explosion, when a

Published in Sergei O. Kuznetsov, Amedeo Napoli, Sebastian Rudolph (Eds.): The 11th International Workshop "What can FCA do for Artificial Intelligence?", FCA4AI 2023, co-located with IJCAI 2023, August 20 2023, Macao, S.A.R. China, Proceedings, pp. 35–46.

vinogradov.d.w@gmail.com (D. Vinogradov)
http://www.ccas.ru (D. Vinogradov)
0000-0001-5761-4706 (D. Vinogradov)

small training context generates exponentially large number of concepts [7]. Another one is the appearance of so-called 'phantom' concepts as accidental similarities between small number of training objects each of which belongs to a different concept with larger extent [8]. It can be argued that the appearance of such hypotheses corresponds to the over-fitting phenomenon. This statement was experimentally confirmed in the master thesis of L.A. Yakimova [9].

To overcome these limitations, the author [10] proposed to use a probabilistic approach. The idea is to generate a random sample of concepts by trajectories of Markov chain making random walks through the concept lattice. We named our approach the VKF method in honor of V.K. Finn and because of the abbreviation of the Russian term "Probabilistic Combinatorial Formal method" to indicate effective processing using probabilistic algorithms and FCA of various combinations of training objects for generating concepts.

Such algorithms are based on the "Close-by-One" operations CbO, for which the part with respect to objects was proposed earlier by S.O. Kuznetsov [11] in the eponymous CbO algorithm for exhaustive generation of all candidates for hypotheses, the number of which in some cases may be exponentially large. Using these operations, the author proposed to generate a polynomial-size random subset of concepts, each element of which corresponds to one trajectory of random walk across the corresponding lattice.

The author [12] has proved that it is sufficient to generate $\frac{n \cdot \ln 2 - \ln \delta}{\varepsilon}$ random concepts in order to correctly predict all the ε-important test objects with the reliability of $1 - \delta$.

Therefore, the single obstacle for polynomial complexity of full scheme of the VKF-method is the absence of polynomial upper bound on the length of trajectories of Markov chain. The main result of this paper is polynomial upper bound on the average length of trajectories of the coupling Markov chain when the training context is dichotomized, i.e., expanded by additional binary attributes that correspond to negations of all original attributes. Such expansion is useful if the absence of an original attribute is allowed to be a part of cause for the target attribute. Previously, only special cases of formal contexts (for instance, Boolean algebra and linear order) were investigated. The new result concerns the general case of arbitrary lattice.

2. Background

2.1. Basic definitions and facts of FCA

Here we recall some basic definitions and facts from Formal Concept Analysis (FCA) [5].

A **(formal) context** is a triple (G, M, I) where G and M are finite sets and $I \subseteq G \times M$. The elements of G and M are called **objects** and **attributes**, respectively. As usual, we write gIm instead of $\langle g, m \rangle \in I$ to denote that object g has attribute m.

For $A \subseteq G$ and $B \subseteq M$, define

$$A' = \{m \in M : \forall g \in A (gIm)\}, \tag{1}$$

$$B' = \{g \in G : \forall m \in B (gIm)\}; \tag{2}$$

so A' is the set of attributes common to all the objects in A and B' is the set of objects possessing all the attributes in B. The maps $(\cdot)' : A \mapsto A'$ and $(\cdot)' : B \mapsto B'$ are called **derivation operators** (also **polars**) of the context (G, M, I).

A **concept** of the context (G, M, I) is defined to be a pair (A, B), where $A \subseteq G$, $B \subseteq M$, $A' = B$, and $B' = A$. The first component A of the concept (A, B) is called the **extent** of the concept, and the second component B is called its **intent**. The set of all concepts of the context (G, M, I) is denoted by $\mathbf{B}(G, M, I)$.

Let (G, M, I) be a context. For concepts (A, B) and (C, D) in $\mathbf{B}(G, M, I)$ we write $(A, B) \leq (C, D)$, if $A \subseteq C$. The relation $\leq$ is a **partial order** on $\mathbf{B}(G, M, I)$.

A subset $A \subseteq G$ is the extent of some concept if and only if $A'' = A$ in which case the unique concept of which A is the extent is (A, A'). Similarly, a subset B of M is the intent of some concept if and only if $B'' = B$ and then the unique concept with intent B is (B', B).

Proposition 1. *Let (G, M, I) be a context. Then $(\boldsymbol{B}(G, M, I), \leq)$ is a lattice with join and meet given by*

$$\bigvee_{j \in J} (A_j, B_j) = ((\bigcup_{j \in J} A_j)'', \bigcap_{j \in J} B_j), \tag{3}$$

$$\bigwedge_{j \in J} (A_j, B_j) = (\bigcap_{j \in J} A_j, (\bigcup_{j \in J} B_j)''); \tag{4}$$

Corollary 1. *For context (G, M, I) the lattice $(\boldsymbol{B}(G, M, I), \leq)$ has (M', M) as the bottom element and (G, G') as the top element. In other words, for all $(A, B) \in \boldsymbol{B}(G, M, I)$ the following inequalities hold:*

$$(M', M) \leq (A, B) \leq (G, G'). \tag{5}$$

For $(A, B) \in \mathbf{B}(G, M, I)$, $g \in G$, and $m \in M$ define

$$CbO((A, B), g) = (A, B) \vee (\{g\}'', \{g\}'), \tag{6}$$

$$CbO((A, B), m) = (A, B) \wedge (\{m\}', \{m\}''), \tag{7}$$

so according to (4) $CbO((A, B), g)$ is equal to $((A \cup \{g\})'', B \cap \{g\}')$ and according to (3) $CbO((A, B), m)$ is equal to $(A \cap \{m\}', (B \cup \{m\})'')$.

The useful properties of introduced operations are summarized in the following Lemmas.

Lemma 1. *Let (G, M, I) be a context, $(A, B) \in \boldsymbol{B}(G, M, I)$, $g \in G$, and $m \in M$. Then*

$$g \in A \Rightarrow CbO((A, B), g) = (A, B), \tag{8}$$

$$m \in B \Rightarrow CbO((A, B), m) = (A, B), \tag{9}$$

$$g \notin A \Rightarrow (A, B) < CbO((A, B), g), \tag{10}$$

$$m \notin B \Rightarrow CbO((A, B), m) < (A, B). \tag{11}$$

Lemma 2. *Let (G, M, I) be a context, $(A, B), (C, D) \in \boldsymbol{B}(G, M, I)$, $g \in G$, and $m \in M$. Then*

$$(A, B) \leq (C, D) \Rightarrow CbO((A, B), g) \leq CbO((C, D), g), \tag{12}$$

$$(A, B) \leq (C, D) \Rightarrow CbO((A, B), m) \leq CbO((C, D), m). \tag{13}$$

2.2. Random walks by coupled Markov chain

To avoid the open problem of calculation of mixing time of general Markov chain we proposed [10] to use the coupled Markov chain for random walks across the concept lattice. The states of this chain are ordered pairs of concepts. The stopping time of the random walk algorithm is the first moment of entering to some ergodic (recurrent) state of the coupled Markov chain. Every ergodic state of the coupled Markov chain is a pair of equal concepts. Denote the set of such states by E.

Data: context (G, M, I), external function $CbO(\ ,\ )$
Result: random concept $(A, B) \in \mathbf{B}(G, M, I)$
$X := G \sqcup M$; $(A, B) := (M', M)$; $(C, D) = (G, G')$;
while $((A \neq C) \vee (B \neq D))$ **do**
  select random element $x \in X$;
  $(A, B) := CbO((A, B), x)$;
  $(C, D) := CbO((C, D), x)$;
end

Algorithm 1: Coupling Markov chain

The algorithm terminates when the upper and lower concepts coincide. The condition on remaining of ordering between two concepts $(A, B) \leq (C, D)$ at any intermediate step of the while loop of Algorithm 1 follows from Lemma 2.

The classical theorem of Markov chain Theory about transient (non-ergodic) states [13] implies almost surely termination of algorithms 1, i.e. finiteness of a trajectory until it enters to some ergodic state with probability 1.

Consider the moment $T_i(E) = \min\{t : X_t \in E, X_0 = s_i\}$ of the first entering to E, starting with an arbitrary transient state $s_i = (\langle A, B\rangle < \langle C, D\rangle) \notin E$.

Theorem 1. *The moment $T_i(E)$ is Markov one for every transient state s_i.*

Proof. We need to prove $\mathbb{P}\left[T_i(E) < \infty \mid X_0 = s_i\right] = 1$.

Use decomposition $\{X_t \in E, X_0 = s_i\} = \bigcup_{n \leq t} U_n(s_i)$, where

$$U_n(s_i) = \{X_n \in E, X_{n-1} \notin E, \ldots, X_1 \notin E, X_0 = s_i\}.$$

Transient States Theorem asserts

$$\lim_{t \to \infty} \mathbb{P}\left[X_t \notin E \mid X_0 = s_i\right] \to 0. \tag{14}$$

Disjointedness of different $U_n(s_i)$ and formula (14) imply

$$\mathbb{P}\{X_t \in E \mid X_0 = s_i\} = \sum_{n \leq t} \mathbb{P}\left[U_n(s_i) \mid X_0 = s_i\right] \to 1,$$

if $t \to \infty$.

Since $U_n(s_i) = \{T_i(E) = n\}$ the σ-additivity leads to needed conclusion. □

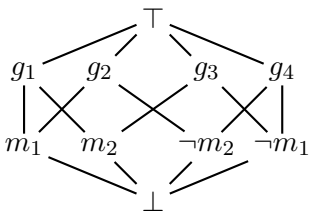
As direct corollary of the theorem 1 we conclude that the termination of algorithm 1 takes place almost surely (i.e. with probability 1).

The goal of current research is to obtain a polynomial upper bound on the average length of trajectories of the coupling Markov chain. In general, it is an open problem. In sequel, we'll provide such bound, when the training context is expanded by additional binary attributes that correspond to negations of all existing attributes (dichotomic expansion).

Example 1. *Dichotomic expansion of the left context is the right one.*

$G \times M$	m_1	m_2		$G \times M^+$	m_1	$\neg m_1$	m_2	$\neg m_2$
g_1	*1*	*1*		g_1	*1*	*0*	*1*	*0*
g_2	*1*	*0*		g_2	*1*	*0*	*0*	*1*
g_3	*0*	*1*		g_3	*0*	*1*	*1*	*0*
g_4	*0*	*0*		g_4	*0*	*1*	*0*	*1*

The expanded context corresponds to the lattice



where $\top = \langle \emptyset, \{m_1, \neg m_1, m_2, \neg m_2, \}\rangle$, $g_j = \langle \{g_j\}, \{g_j\}'\rangle$, $m_j = \langle \{m_j\}', \{m_j\}\rangle$, $\neg m_j = \langle \{\neg m_j\}', \{\neg m_j\}\rangle$, *and* $\bot = \langle \{g_1, g_2, g_3, g_4\}, \emptyset\rangle$.

The coupled Markov chain starts with state $(\bot \leq \top)$. *A trajectory of the random walk depends on random choices from* $G \sqcup M^+$.

Consider an example of such trajectory. Assume that g_1 *is selected at the 1st step, then the chain goes to state* $(\bot \leq g_1)$. *The choice of* g_2 *at the 2nd step leads to* $(\bot \leq m_1)$. *If the chain selects* $\neg m_1$ *at the 3rd step, then the state becomes* $(\neg m_1 \leq \top)$. *The choice of* g_4 *at the 4th step leads to* $(\neg m_1 \leq g_4)$. *After selection of* g_3 *at the 5th step the trajectory goes to ergodic state* $(\neg m_1 \leq \neg m_1)$, *and algorithm 1 stops.*

3. Technical Tools

In [14] the author developed a useful tool to estimate the average length of trajectories of coupling Markov chain through recurrence relations.

Lemma 3.

$$\mathbb{E}\left[T_i(E)\right] = 1 + \sum_{s_j \notin E} \mathbb{E}\left[T_j(E)\right] \cdot \mathbb{P}\left[X_1 = s_j | X_0 = s_i\right]$$

for every $s_i \notin E$.

Proof. Additivity of the average gives

$$\mathbb{E}\left[T_i(E)\right] = \sum_{n=1}^{\infty} n \cdot \mathbb{P}\left[U_n(s_i) | X_0 = s_i\right], \tag{15}$$

where $U_n(s_i) = \{X_n \in E, X_{n-1} \notin E, \ldots, X_1 \notin E, X_0 = s_i\}$.

Then

$$\begin{aligned}
\mathbb{E}\left[T_i(E)\right] = \sum_{n=1}^{\infty} n \cdot \mathbb{P}\left[U_n(s_i) | X_0 = s_i\right] = \\
= \sum_{n=1}^{\infty} \mathbb{P}\left[U_n(s_i) | X_0 = s_i\right] + \sum_{n=2}^{\infty} (n-1) \cdot \mathbb{P}\left[U_n(s_i) | X_0 = s_i\right] = \\
= 1 + \sum_{k=1}^{\infty} k \cdot \mathbb{P}\left[X_{k+1} \in E, X_k \notin E, \ldots, X_1 \notin E | X_0 = s_i\right] = 1 + \\
+ \sum_{s_j \notin E} \sum_{k=1}^{\infty} k \cdot \mathbb{P}\left[X_{k+1} \in E, X_k \notin E, \ldots, X_2 \notin E | X_1 = s_j\right] \cdot \mathbb{P}\left[X_1 = s_j | X_0 = s_i\right] = \\
= 1 + \sum_{s_j \notin E} \mathbb{E}\left[T_j(E)\right] \cdot \mathbb{P}\left[X_1 = s_j | X_0 = s_i\right].
\end{aligned}$$

Here we sequentially use identity (15), the Markov property for moment $T_i(E)$ (theorem 1) and the Law of Total Probability. □

This easily results in an upper bound of the order $O(n \cdot \ln n)$ on the average length of trajectories of algorithm 1 for n-dimensional Boolean algebra case.

A more striking result from [14] concerns the average trajectory length of the algorithm 1 for linear orders. Here the upper bound of 4 on the average length does not depend on the number of elements of the linear order.

Example 2. *Apply lemma 3 to the lattice from example 1.*

This lattice allows us to define a distance between ordered candidates (i.e. components of a state). State $s_0 = (\bot \leq \top)$ *has distance 3. States* $s_1 = (\bot \leq g_1), \ldots, s_4 = (\bot \leq g_4)$, $s_5 = (m_1 \leq \top)$, $s_6 = (m_2 \leq \top)$, $s_7 = (\neg m_2 \leq \top)$, $s_8 = (\neg m_1 \leq \top)$ *have distance 2. States with distance 1 are divided into 2 groups (external and internal ones). External states are* $s_9 = (\bot \leq m_1)$, $s_{10} = (\bot \leq m_2)$, $s_{11} = (\bot \leq \neg m_2)$, $s_{12} = (\bot \leq \neg m_1)$, *and* $s_{13} = (g_1 \leq \top), \ldots, s_{16} = (g_4 \leq \top)$. *Internal states correspond to edges* $s_{17} = (m_1 \leq g_1), \ldots, s_{24} = (\neg m_1 \leq g_4)$. *The rest states are ergodic ones (with distance 0).*

Denote the length of trajectory starting from state s_0 *by* T_3. *The states with distance 2 determine a random walk of length* T_2. *External states initiate trajectories of length* T_1^E, *and internal ones start trajectories of length* T_1^M.

Lemma 3 leads to $\mathbb{E}T_3 = 1 + \mathbb{E}T_2$ *and system of equations*

$$\begin{cases}
\mathbb{E}T_2 = 1 + \frac{3}{8}\mathbb{E}T_2 + \frac{2}{8}\mathbb{E}T_1^E + \frac{2}{8}\mathbb{E}T_1^M \\
\mathbb{E}T_1^E = 1 + \frac{1}{8}\mathbb{E}T_2 + \frac{2}{8}\mathbb{E}T_1^E + \frac{2}{8}\mathbb{E}T_1^M \\
\mathbb{E}T_1^M = 1 + \frac{2}{8}\mathbb{E}T_1^E + \frac{2}{8}\mathbb{E}T_1^M
\end{cases} \tag{16}$$

The solution $\mathbb{E}T_2 = \frac{288}{72} = 4$ *of the system (16) leads to the value* $\mathbb{E}T_3 = 1 + 4 = 5$ *of the average length of trajectory of algorithm 1 for the context considered in example 1.*

Now we extend component-wise the CbO operations to states of coupling Markov chain

$$CbO((\langle A, B\rangle \le \langle C, D\rangle), g) = (CbO(\langle A, B\rangle, g) \le CbO(\langle C, D\rangle, g))$$

and

$$CbO((\langle A, B\rangle \le \langle C, D\rangle), m) = (CbO(\langle A, B\rangle, m) \le CbO(\langle C, D\rangle, m)).$$

Then we define a (partial) order between states $s_i = (\langle A_i, B_i\rangle \le \langle C_i, D_i\rangle)$ and $s_j = (\langle A_j, B_j\rangle \le \langle C_j, D_j\rangle)$ of coupling Markov chain as following

$$s_j \leqslant s_i \Leftrightarrow \langle A_i, B_i\rangle \le \langle A_j, B_j\rangle \le \langle C_j, D_j\rangle \le \langle C_i, D_i\rangle. \tag{17}$$

Lemma 2 easily implies

Lemma 4. *For any ordered pair of states* $s_j \leqslant s_i$*, any* $g \in G$*, and any* $m \in M$ $CbO(s_j, g) \leqslant CbO(s_i, g)$ *and* $CbO(s_j, m) \leqslant CbO(s_i, m)$ *hold.*

We denote the number of training objects by $k = |G|$ and the number of attributes by $n = |M|$.

Lemma 5. $\mathbb{E}T_j(E) \le \mathbb{E}T_i(E)$ *for any ordered pair of transient states* $s_j \leqslant s_i$ *of coupling Markov chain.*

Proof. Define coupled random walk of ordered pair of states $X_t \leqslant Y_t$ as following:

$$\mathbb{P}\left[X_1 = s'_j, Y_1 = s'_i \mid X_0 = s_j, Y_0 = s_i\right] = \\ = \begin{cases} \frac{l}{n+k}, & l = |\{g \in G : s'_j = CbO(s_j, g), s'_i = CbO(s_i, g)\}| + \\ & +|\{m \in M : s'_j = CbO(s_j, m), s'_i = CbO(s_i, m)\}| \\ 0, & \neg\exists g \in G \left[s'_j = CbO(s_j, g), s'_i = CbO(s_i, g)\right] \,\& \\ & \&\neg\exists m \in M \left[s'_j = CbO(s_j, m), s'_i = CbO(s_i, m)\right] \end{cases}.$$

Lemma 4 implies $\mathbb{P}[X_1 \leqslant Y_1 \mid X_0 \leqslant Y_0] = 1$.

Since $\langle A_i, B_i\rangle = \langle C_i, D_i\rangle$ for $\langle A_i, B_i\rangle \le \langle A_j, B_j\rangle \le \langle C_j, D_j\rangle \le \langle C_i, D_i\rangle$ implies $\langle A_i, B_i\rangle = \langle A_j, B_j\rangle = \langle C_j, D_j\rangle = \langle C_i, D_i\rangle$, then by definitions it follows that

$$\mathbb{P}[X_t = Y_t \in E \mid X_0 = s_j \leqslant Y_0 = s_i] \ge \mathbb{P}[Y_t \in E \mid X_0 = s_j \leqslant Y_0 = s_i]. \tag{18}$$

Recall that for an integer-valued random variable Z, the equality $\mathbb{E}Z = \sum_{t=0}^{\infty} \mathbb{P}[Z > t]$ is fulfilled. Now $X_t \notin E \Leftrightarrow T_i(E) > t$ and $Y_t \notin E \Leftrightarrow T_j(E) > t$.

Therefore, equation (18) implies

$$\mathbb{P}[T_j(E) > t \mid X_0 = s_j, Y_0 = s_i] \le \mathbb{P}[T_i(E) > t \mid X_0 = s_j, Y_0 = s_i],$$

and the summation over t leads to the required result. □

4. Main result

In the following we'll assume $G' = \emptyset$. This is easily achieved by eliminating all the attributes common to all training objects.

Let's dichotomize the context, i.e., enrich the set of attributes by introducing an attribute for the negation $\neg m_j$ of every binary attributes $m_j \in M$. This construction often has a useful meaning: we want the absence of a attribute to be a new attribute, i.e., we propose dichotomic scaling of the context (according to [5]).

The enriched set of attributes will be denoted by M^+, and we denote its power by $2n = |M^+|$. Usually $2n \ll k = |G|$, which we will assume in the future. Enrich the training context to $I \subseteq G \times M^+$ by the rule:

$$gI\neg m_j \Leftrightarrow \neg(gIm_j).$$

Divide all transient states into 2 groups:

$$V = \{s = (\langle A, B\rangle < \langle C, D\rangle) : \exists m \in M^+ \, [m \in B]\} \tag{19}$$

and

$$W = \{s = (\langle A, B\rangle < \langle C, D\rangle) : \forall m \in M^+ \, [m \notin B]\}. \tag{20}$$

It is clear that the state $s_0 = (\bot < \top) \in W$. By lemma 5 for any $s_j \in W, \mathbb{E}T_j(E) \leq \mathbb{E}T_0(E)$.

By the definition of the set V and the lemma 5 for any $s_j \in V$ we have $\mathbb{E}T_j(E) \leq \mathbb{E}T_i(E)$, where $s_i = (\langle \{m\}', \{m\}''\rangle < \top) \in V$ for any $m \in B$ with $s_j = \langle A, B\rangle$.

Let's introduce an integer-valued random variable Z taking the value q on the event $\{X_q = (\bot = \bot), X_{q-1} \notin V, \ldots, X_1 \notin V, X_0 = s_0\}$, which determines the minimum number of steps of the algorithm 1 by states from $X_t \in W$ until we get $X_q = (\bot = \bot)$.

Lemma 6.

$$\mathbb{E}Z = \sum_{l=1}^{\infty} \mathbb{P}\left[Z \geq l\right] \leq (k + 2n) \cdot \left(\ln(2n) + \frac{1}{1 - e^{-1}}\right)$$

for context $I \subseteq G \times M^+$ with $2n = |M^+| \leq k = |G|$.

Proof. We divide the summands into disjoint subsets of $I_0 \sqcup \bigsqcup_{r=1}^{\infty} I_r$, where $I_0 = \{1 \leq l < (k + 2n) \cdot \ln(2n)\}$ and

$$I_r = \{(k + 2n) \cdot (\ln(2n) + r - 1) \leq l < (k + 2n) \cdot (\ln(2n) + r)\}.$$

It is clear that $\sum_{l=1}^{(k+2n)\cdot\ln(2n)-1} \mathbb{P}\left[Z \geq l\right] \leq (k + 2n) \cdot \ln(2n)$.

In order for the event $Z \geq l$ to occur, it is necessary that at least one attribute (out of $2n$) is selected, so that no example in the series of length l is selected in which this attribute is not present. Therefore, by Boole's inequality

$$\mathbb{P}\left[Z > l\right] \leq 2n \cdot \left(1 - \frac{1}{k + 2n}\right)^l.$$

For I_r

$$\sum_{l=(k+2n)\cdot(\ln(n)+r-1)}^{(k+2n)\cdot(\ln(2n)+r)-1} \mathbb{P}\left[Z > l\right] \leq \sum_{l=(k+2n)\cdot(\ln(2n)+r-1)}^{(k+2n)\cdot(\ln(2n)+r)-1} k \cdot \left(1 - \frac{1}{k+2n}\right)^{l} \leq$$

$$\leq (k+2n) \cdot 2n \cdot \left(1 - \frac{1}{k+2n}\right)^{(k+2n)\cdot(\ln(2n)+r-1)} \leq$$

$$\leq (k+2n) \cdot e^{\ln 2n} \cdot e^{-(\ln(2n)+r-1)} = (k+2n) \cdot e^{-r+1}.$$

The summation over r gives

$$\sum_{l=(k+2n)\cdot\ln(2n)}^{\infty} \mathbb{P}\left[Z \geq l\right] \leq (k+2n) \cdot \sum_{r=1}^{\infty} e^{-r+1} = \frac{k+2n}{1-e^{-1}}.$$

□

Let's denote the upper bound from lemma 6 by $tail$.

We consider disjoint events

$$H_l(s_j) = \{X_l = s_j \in V, X_{l-1} \notin V, \ldots, X_1 \notin V, X_0 = s_0\}. \tag{21}$$

We denote event $\{X_{t+l} \in E, X_{t+l-1} \notin E, \ldots, X_{l+1} \notin E\} \cap H_l(s_j)$ by $G_{t,l}(s_j)$, and the union $\bigsqcup_{s_j \in V} G_{t,l}(s_j)$ by $U_{t,l}$.

It is clear that we have a decomposition of the event into disjoint parts

$$\{X_{t+l} \in E, X_{t+l-1} \notin E, \ldots, X_0 = s_0\} =$$

$$= \bigsqcup_{s_j \in V} (\{X_{t+l} \in E, X_{t+l-1} \notin E, \ldots, X_{l+1} \notin E\} \cap H_l(s_j)) \sqcup$$

$$\sqcup \{X_{t+l} = (\bot = \bot), X_{t+l-1} \notin V, \ldots, X_1 \notin V, X_0 = s_0\}.$$

We need

$$\mathbb{E}T_0(E) = \sum_{m=1}^{\infty} m \cdot \mathbb{P}\left[X_m \in E, X_{m-1} \notin E, \ldots, X_1 \notin E \mid X_0 = s_0\right]. \tag{22}$$

It is clear that $\mathbb{E}T_0(E) = \mathbb{E}T_0'(E) + \mathbb{E}Z$, where $T_0'(E)$ is restriction of $T_0(E)$ on $\bigsqcup_{t=1}^{\infty} \bigsqcup_{l=1}^{\infty} G_{t,l}$.

Theorem 2. *For the dichotomized (enriched) training context $I \subseteq G \times M^+$ with $2n = |M^+| \leq k = |G|$ the upper bound on the average length of trajectories of algorithm 1 is*

$$\mathbb{E}T_0 \leq \frac{(k+2n)(k^2+k(2n+1)+4n^2+2n)}{2n(k^2+k+2n)} + \frac{(k+1)(k+2n)}{k^2+k+2n} tail.$$

Proof. Let's denote $R = \sum_{l=1}^{n} \frac{1}{k+2n} (T_{f_l} + T_{\neg f_l})$, where $T_{f_l} = T_i(E)$ for $s_j = (\langle \{f_l\}', \{f_l\}'' \rangle < \top)$, and similarly for $T_{\neg f_l}$.

Then Markov property implies

$$\mathbb{E}T_0'(E) = \sum_{t=1}^{\infty}\sum_{l=1}^{\infty}(t+l)\cdot\mathbb{P}U_{t,l} =$$

$$= \sum_{t=1}^{\infty} t \cdot \sum_{s_j \in V} \mathbb{P}\left[X_t \in E, X_{t-1} \notin E, \ldots, X_1 \notin E \mid X_0 = s_j\right] \cdot \sum_{l=1}^{\infty} \mathbb{P}\left[H_l(s_j)\right] +$$

$$+ \sum_{l=1}^{\infty} l \cdot \sum_{s_j \in V} \mathbb{P}\left[H_l(s_j)\right] \cdot \sum_{t=1}^{\infty} \mathbb{P}\left[X_t \in E, X_{t-1} \notin E, \ldots, X_1 \notin E \mid X_0 = s_j\right] \leq$$

$$\leq \sum_{s_j \in V} \mathbb{E}T_j(E) \cdot \mathbb{P}\left[X_1 = s_j \mid X_0 = s_0\right] + \sum_{s_j \in V}\sum_{l=1}^{\infty} l \cdot \mathbb{P}\left[H_l(s_j)\right] \leq$$

$$\leq \mathbb{E}R + \frac{k+2n}{2n},$$

where the last term is the average of geometrically distributed random variable of the time before first selection of some attribute.

The Law of Total Probability and lemma 5 imply

$$\mathbb{E}T_{f_l} \leq 1 + \sum_{i=1}^{n} \frac{1}{k+2n}\left(\mathbb{E}T_{f_i} + \mathbb{E}T_{\bar{f}_i}\right) - \frac{1}{k+2n} \cdot \mathbb{E}T_{\bar{f}_l} + \frac{k}{k+2n}\mathbb{E}T_0(E).$$

Therefore,

$$\mathbb{E}R \leq \frac{2n}{k+2n}\left[1 + \mathbb{E}R + \frac{k}{k+2n}\mathbb{E}T_0(E)\right] - \frac{1}{k+2n}\mathbb{E}R.$$

Hence,

$$\frac{k+1}{k+2n}\mathbb{E}R \leq \frac{2n}{k+2n} + \frac{2nk}{(k+2n)^2}\mathbb{E}T_0(E).$$

Substitute $\mathbb{E}R \leq \frac{2n}{k+1} + \frac{2nk}{(k+1)(k+2n)}\mathbb{E}T_0(E)$ into

$$\mathbb{E}T_0(E) \leq \mathbb{E}R + \frac{k+2n}{2n} + tail,$$

and obtain

$$\frac{k^2+k+2n}{(k+1)(k+2n)}\mathbb{E}T_0(E) \leq \frac{2n}{k+1} + \frac{k+2n}{2n} + tail,$$

which leads to the required result. □

5. Conclusion

In this article we have presented a significant advancement in solving the open problem of the VKF method about finding a polynomial upper bound on the average length of trajectories of a coupling Markov chain - the average time of computation by the probabilistic algorithm 1, which generates concepts of the training context for knowledge extraction. Only special cases,

such as Boolean algebra and linear order, were investigated earlier. The important step is based on the dichotomic scaling of a training context.

Combining the new result with the previously obtained polynomial lower bound on the sufficient number of concepts, we obtain a fully polynomial scheme for extracting knowledge using a binary similarity operation implemented in the VKF-method.

Experimental studies of author's PhD student L.A. Yakimova demonstrate that probabilistic approach to FCA-based knowledge extraction (in combination with "Counterexample Forbidding Condition") is practically not subject to the phenomenon of over-fitting (through generation of 'phantom' candidates), unlike the classical JSM-method.

Acknowledgments

The author thanks his colleagues from Dorodnicyn Computing Center of Federal Research Center "Computer Science and Control" of Russian Academy of Sciences for support and useful discussions. The author is grateful to his PhD student Lyudmila A. Yakimova for long-term cooperation that stimulated the described research.

References

[1] V. K. Finn, J.S.Mill's inductive methods in artificial intelligence systems I, Scientific and Technical Information Processing 38 (2011) 385–402. doi:10.3103/S0147688211060037.

[2] V. K. Finn, J.S.Mill's inductive methods in artificial intelligence systems II, Scientific and Technical Information Processing 39 (2012) 241–260. doi:10.3103/S0147688212050036.

[3] J. S. Mill, A System of Logic: Ratiocinative and Inductive, John W. Parker, London, 1843.

[4] S. O. Kuznetsov, Machine learning on the basis of formal concept analysis, Automation and Remote Control 62 (2001) 1543–1564. doi:10.1023/A:1012435612567.

[5] B. Ganter, R. Wille, Formal Concept Analysis: Mathematical Foundations, Springer, Berlin, Heidelberg, 1999. doi:10.1007/978-3-642-59830-2.

[6] S. O. Kuznetsov, S. A. Obiedkov, Algorithms for the construction of concept lattices and their diagram graphs, in: L. D. Raedt, A. Siebes (Eds.), Proceedings of the 5th Conference on Principles of Data Mining and Knowledge Discovery, volume 2168 of *Lecture Notes in Artificial Intelligence*, Springer, Berlin, Heidelberg, 2001, pp. 289–300. doi:10.1007/3-540-44794-6.

[7] D. V. Vinogradov, Existence of large sublattices isomorphic to boolean algebra in a candidate lattice, Autom. Doc. Math. Linguist. 57 (2023) 101–103. doi:10.3103/S0005105523020097.

[8] D. V. Vinogradov, Accidental formal concepts in the presence of counterexamples, in: S. O. Kuznetsov, B. W. Watson (Eds.), Proceedings of International Workshop on Formal Concept Analysis for Knowledge Discovery, volume 1921 of *CEUR Workshop Proceedings*, HSE, Moscow, Russia, 2017, pp. 104–112.

[9] L. A. Yakimova, Experimental investigation of behaviour of solvers based on binary similarity operation, Master's thesis, Russian State University for Humanities, Moscow, Russia, 2020. In Russian.

[10] D. V. Vinogradov, VKF-method of hypotheses generation, in: Proceedings of the 3rd International Conference on Analysis of Images, Social Networks and Texts (AIST'2014), volume 436 of *Communications in Computer and Information Science*, 2014, pp. 237–248. doi:10.1007/978-3-319-12580-0_25.

[11] S. O. Kuznetsov, A fast algorithm for computing all intersections of objects from an arbitrary semilattice, Nauch.-Tekh. Inf. Ser.2 27 (1993) 17–20.

[12] D. V. Vinogradov, Algebraic machine learning: Emphasis on efficiency, Automation and Remote Control 83 (2022) 831–846. doi:10.1134/S0005117922060029.

[13] J. G. Kemeny, J. L. Snell, Finite Markov Chains, Undergraduate Texts in Mathematics, 1 ed., Springer, New York, 1976. Originally published by Van Nostrand Publishing Company, 1960.

[14] D. V. Vinogradov, Markov chains, law of total probability, and recurrence relations, Autom. Doc. Math. Linguist. 57 (2023) 68–72. doi:10.3103/S0005105523010090.