Small Overfitting Probability in Minimization of Empirical Risk for FCA-based Machine Learning

Dmitry V. Vinogradov^{1,*}

¹Dorodnicyn Computing Center, Federal Research Center for Computer Science and Control, Russian Academy of Sciences, Moscow 119333, Russia

Abstract

Main result of the paper provides a small upper bound on a probability of overfitting in FCA-based Machine Learning in the simplest case of Boolean algebra without counter-examples. This Machine Learning approach uses a set of randomly generated formal concepts to predict test examples. The well-known Vapnik–Chervonenkis' technique of empirical risk minimization determines a number of generated concepts. Estimations of Mixture and Stopping times of several probabilistic algorithms based on Markov chains leads to a plausible assumption on the uniform independent distribution of elements of Boolean algebra. In this case the main theorem proves that the probability of overfitting in fixed fraction of test examples tends to zero faster than exponent when the number of attributes goes to infinity.

Keywords

FCA, Markov chain, empirical risk, overfitting, Boolean algebra

1. Introduction

Formal Concept Analysis (FCA) [1] is a popular means of data analysis in case of small samples. However an applicability of FCA to Big Data has several obstacles:

- There will be an exponentially large number of hypotheses with respect to a size of an formal context in the worst case (for instance, the case of Boolean algebra, see next section).
- Many problems of FCA [4] belong to famous classes of *NP* and *#P*-complete problems of computational complexity.
- There is a positive probability of "accidental" concepts appearance that corresponds to the overfitting phenomenon [9].

The paper [8] introduces Markov chain approach to probabilistic generation of formal concepts to construct a Machine Learning system based on FCA (FCAML). Recently FCAML-system

Published in Sergei O. Kuznetsov, Amedeo Napoli, Sebastian Rudolph (Eds.): The 10th International Workshop "What can FCA do for Artificial Intelligence?", FCA4AI 2022, co-located with IJCAI-ECAI 2022, July 23 2022, Vienna, Austria, Proceedings, pp. 41–49.

^{*}Corresponding author.

[☆] vinogradov.d.w@gmail.com (D. V. Vinogradov)

http://ccas.ru (D. V. Vinogradov)

D 0000-0001-5761-4706 (D. V. Vinogradov)

^{© 02022} Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

applies the coupling Markov chain to generate a random sample of concepts. Each run of this chain terminates with probability 1. Early the system uses a monotonic Markov chain that corresponds to the famous Lazy Random Walk in the case of Boolean algebra. The paper [10] discuses Induction procedure for generalization of training examples into hypotheses about causes of the property under research with counter-examples forbidding. Finally the system predicts a target class of each test example by Analogy reasoning.

The main result of paper [10] gives a sufficient number of hypotheses to predict a target class with given level of confidence. The framework is dual to the famous one of V.N. Vapnik and A.Ya. Chervonenkis with respect to their dimension of a classifiers class.

However V.N. Vapnik and A.Ya. Chervonenkis also developed another approach to choose a good hypothesis. Best one must minimize an empirical risk (a fraction of wrongly predicted training examples). The main problem in this approach is to estimate a fraction of test examples with prediction error. We provide a partial answer on this task in the simplest case of Boolean algebra without counter-examples.

2. Background

2.1. Basic definitions and facts

Here we recall some basic definitions and facts of Machine Learning based on Formal Concept Analysis (FCAML) in the particular case of Boolean algebra. Most general situation is considered in [10]. Book [1] is the best source of information about Formal Concept Analysis itself.

The smallest (formal) context for *n*-dimensional Boolean algebra is a triple (G, M, \neq) , where $G = \{g_1, \ldots, g_n\}$ is a set of coatoms (objects), $M = \{m_1, \ldots, m_n\}$ is a set of binary attributes, and $\neq \subseteq G \times M$ is relation defined as $g_i \neq m_j \Leftrightarrow i \neq j$.

A **concept** of the Boolean algebra context (G, M, \neq) is defined to be a pair (A, B), where $A \subseteq G, B \subseteq M, A = \{g_i \in G : \forall m_j \in B [i \neq j]\}$. 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**. It clear that the lattice of concepts coincides with Boolean algebra of all the subsets $B \subseteq M$. We consider the partial order on it dual to set inclusions.

This observation proves truth of the first obstacle of applicability of FCA to Big Data from Introduction. When formal context occupies n^2 bites only, the full description of *n*-dimensional Boolean algebra requires $n \cdot 2^n$ bites of memory.

Proposition 1. For context (G, M, \neq) corresponding Boolean algebra of concepts has (\emptyset, M) as the bottom element and (G, \emptyset) as the top element. In other words, for every the concept (A, B) the following inequalities hold:

$$(\emptyset, M) \le (A, B) \le (G, \emptyset). \tag{1}$$

Definition 1. For a concept (A, B), $g_i \in G$, and $m_j \in M$ define

$$CbO((A,B),g_i) = \begin{cases} (A,B) & g_i \in A\\ (A \cup \{g_i\}, B \setminus \{m_i\}) & g_i \notin A \end{cases}$$
(2)

$$CbO((A,B),m_j) = \begin{cases} (A,B) & m_j \in B\\ (A \setminus \{g_j\}, B \cup \{m_j\}) & m_j \notin B \end{cases}$$
(3)

We call these operations CbO because the first one is used in Close-by-One (CbO) Algorithm to generate all the formal concepts of an arbitrary context, see [3] for details.

Monotonicity property of introduced operations are summarized in the following Lemma.

Lemma 1. Let (G, M, \neq) be a context, (A, B) and (C, D) be concepts for it, $g \in G$, and $m \in M$. Then

$$(A,B) \le (C,D) \Rightarrow CbO((A,B),g) \le CbO((C,D),g),\tag{4}$$

$$(A,B) \le (C,D) \Rightarrow CbO((A,B),m) \le CbO((C,D),m).$$
(5)

Initially the system used the monotonic Markov chain algorithm as a core of probabilistic approach to Machine Learning based on FCA.

Data: context (G, M, \neq) **Result:** random concept (A, B) $V := G \sqcup M; (A, B) := (\emptyset, M);$ for (i := 0; i < T; i + +) do select random element $v \in V$; (A,B) := CbO((A,B),v);

end

Algorithm 1: Monotonic Markov chain

The main difficulty with the monotonic Markov chain in general case is an absence of a good estimation on length T of its trajectory to achieve approximately stationary distribution. However the case of Boolean algebra was investigated successfully by the author. Next subsection contains the key results about it.

Then the coupling Markov chain algorithm described below appears, where there exists the natural stopping moment. Now it is a working horse for our approach.

Data: context (G, M, \neq) **Result:** random concept (A, B) $V := G \sqcup M; (A, B) := (\emptyset, M); (C, D) = (G, \emptyset);$ while $(A \neq C)$ do select random element $v \in V$; (A, B) := CbO((A, B), v);(C, D) := CbO((C, D), v);end

Algorithm 2: Coupling Markov chain

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

Now we represent Machine Learning based on FCA (FCAML-method) for our setting (Boolean algebra without counter-examples). See seminal paper[10] for description of the general scheme of FCAML-method.

The reader can learn the classical deterministic FCA-based approach to Machine Learning from Kuznetsov [5]. Our technique uses probabilistic Algorithm 2 for computing a random subset of concepts.

As usual, there are two sets of objects called the training $G = \{g_1, \ldots, g_n\}$ and test $X = \{o_1, \ldots, o_n\}$ samples, respectively. Set X contains examples to predict the target class (so-called test objects).

From the training samples the program generates a formal context (G, M, \neq) , where $M = \{m_1, \ldots, m_n\}$. After that the program applies the coupling Markov chain Algorithm 2 to generate a given number N of random concepts (A, B).

Data: number N of concepts to generate **Result:** random sample S of formal concepts **while** (i < N) **do** Generate concept (A, B) by Algorithm 2; $S := S \cup \{(A, B)\};$ i := i + 1;end

Algorithm 3: Inductive generalization

FCAML-method replaces a time-consuming deterministic algorithm (for instance, "Close-by-One") for generation of all concepts by the probabilistic one to randomly generate the prescribed number of concepts. The goal of Markov chain approach is to select a random sample of formal concepts without computation of the (possibly exponential size) set of all the concepts.

How to select number N of concepts for the coupling Markov chain? There are 2 different approaches, both based on ideas of V.N. Vapnik and A.Ya. Chervonenkis. The approach promoted by K.V. Vorontsov is the empirical risk minimization.

Data: random sample S of concepts

Result: empirical risk valueG :=training examples; k := 0;for $(g \in G)$ dofor $(\langle A, B \rangle \in S)$ doif $(B \subseteq \{g\}')$ then $| k := k + \frac{1}{n};$ endbreak;end

end

Algorithm 4: Calculation of empirical risk

In the Boolean algebra case it is possible select sufficiently large N to make the empirical rick equals to 0.

With permutation, we can assume without reducing generality that the first n objects were included in the training sample, and the last n objects form the test sample.

Let N of FCA Machine Learning hypotheses be generated using a coupling Markov chain from a training sample for Boolean algebra, where N chosen sufficiently large to obtain $\eta = 0$.

Stationary distribution uniformity on hypotheses allows to construct hypothesis $h_j = (A_j, B_j)$ (where $1 \le j \le N$) from independent Bernoulli sequence $E_j = (\epsilon_{j,1}, \ldots, \epsilon_{j,n})$ as $A_j = \{g_i : \epsilon_{j,i} = 0\}$ and $B_j = \{f_i : \epsilon_{j,i} = 1\}$.

Finally, the FCAML program predicts the target class of test examples and computes tests

risk.

Consider set F of binary features $F = \{f_1, \ldots, f_n\}$. For each $1 \le i \le n$ introduce test example g_{n+i} with the intent $\{g_{n+i}\}' = \{f_j \in F : j \ne i\}$. Independent Bernoulli sequence $E_j = (\epsilon_{j,1}, \ldots, \epsilon_{j,n})$ determines the corresponding hypothesis $h_j = (A_j, B_j)$, where $A_j = \{o_i : \epsilon_{j,i} = 0\}$ and $B_j = \{f_i : \epsilon_{j,i} = 1\}$. **Data:** random sample $S = \{(A_1, B_1), \ldots, (A_N, B_N)\}$ of concepts **Result:** number of erroneous predicted test examples X :=test examples; r = 0; **for** $(g \in X)$ **do** | **for** (j := 0; j < N; j + +) **do** | **if** $(B_j \subseteq \{g\}')$ **then** | r := r + 1;**end** break; **end**

end

Algorithm 5: Calculation of fraction of erroneous predictions

2.2. Approximate Uniformity of Random Subsets

Algorithm 1 in the case of Boolean algebra coincides with famous Lazy Random Walk on Boolean hypercube.

Lemma 2. Stationary distribution π of Lazy Random Walk is uniform.

The simplest proof of Lemma 2 uses reversibility of the corresponding Markov chain with Balance equations with respect to uniform distribution π .

Definition 2. Total variation distance between distributions $\mu = \langle \mu_i : g_i \in G \rangle$ and $\pi = \langle \pi_i : g_i \in G \rangle$ on finite space G is defined as the half of L_1 -metric, i.e. $\|\mu - \pi\|_{TV} = \frac{1}{2} \cdot \sum_{g_i \in G} |\mu_i - \pi_i|$.

Lemma 3. $\|\mu - \pi\|_{TV} = \max_{A \subseteq G} |\mu(A) - \pi(A)|.$

This Lemma is a direct consequence of Definition 2.

Proposition 2. For Lazy Random Walk let

$$\mu(0) = \mathbb{P}[X_{t+1} = g_i \mid X_t = g_i] = \frac{1}{2},$$

$$\mu(e_j) = \mathbb{P}[X_{t+1} = (g_i \oplus e_j) \mid X_t = g_i] = \frac{1}{2n}$$

and $\mu = 0$ otherwise, and let π be the uniform distribution. Then

$$\left(\|\mu^{*t} - \pi\|_{TV}\right)^2 \le \frac{1}{4} \cdot \left(e^{e^{-c}} - 1\right).$$

holds for $t \geq \frac{1}{2} \cdot n \cdot (\log n + c)$.

This proposition is analogue of result of Diakonis [2] and it was proved by the author during his research on monotonic Markov chain (Algorithm 1).

Comparison of Algorithms 1 and 2 gives assertion that the lower component of the coupling Markov chain coincides with a state of the monotonic Markov chain.

The next two propositions estimate mean length $\mathbb{E}T$ of trajectory of coupling Markov chain (Algorithm 2) and proves the strong concentration of trajectory length T around the mean $\mathbb{E}T$. The author proved them during research on coupling Markov chain (Algorithm 2).

Proposition 3. For *n*-dimensional Boolean algebra

$$\mathbb{E}T = \sum_{j=1}^{n} \frac{n}{j} \approx n \cdot \ln(n) + n \cdot \gamma + \frac{1}{2}.$$
(6)

Proposition 4. For *n*-dimensional Boolean algebra

$$\mathbb{P}\left[T \ge (1+\varepsilon) \cdot n \cdot \ln(n)\right] \to 0,\tag{7}$$

when $n \to \infty$ for any $\varepsilon > 0$.

Statements of Propositions 2 and 3 and Lemma 4 imply the assertion that outputs of Algorithm 2 are approximately uniformly distributed.

Since each trajectory depends only on (pseudo-)random number generator, these outputs are independent.

But random subsets of binary attributes can be generated by Bernoulli sequences. It provides possibility of direct probabilistic computations. In the next section these considerations lead to main result of the paper.

3. Main result

Any set of hypotheses about causes of the target property can be considered as a classifier: if a test example includes at least one hypothesis, then the classifier will predict the target class positively; if none of the hypothetical reasons is embedded in a test example, then this example is classified negatively.

The method of minimizing empirical risk proposed by V.N. Vapnik and A.Ya. Chervonenkis [7] consists in choosing algorithms for which the classification of training examples contains the minimum number of errors (empirical, or observed, risk). In our case, there will always be classifiers (sets of hypotheses) whose empirical risk is zero. We restrict ourselves to these situations only. On the other hand, a risk of making a mistake in predicting test examples remains.

Following K.V. Vorontsov[11], we will randomly divide objects into two groups: training and test examples. For simplicity, let's assume that the number of objects is even, and the splitting is done in half. This assumption does not reduce generality, since the mean binomial coefficient is the largest one.

Let's denote the empirical risk by η and introduce the prediction risk as a fraction of $\theta = r/n$ incorrectly predicted test examples. We are interested in the probability of $\mathbb{P}[\eta = 0, \theta = \delta]$ when objects are evenly divided into training and test samples in half. Since the probabilities are equal for each partition, we can assume without reducing generality that the first n objects were included in the training sample, and the last n objects form the test sample.

Let N of FCAML hypotheses be generated using a coupling Markov chain from a training sample for Boolean algebra. If the trajectories are chosen long enough, then the distribution of hypotheses will be (almost) uniform and independent. Uniformity follows from the property of fast mixing to a uniform stationary distribution, and independence follows from the independence of the Markov chain trajectories generating the FCAML hypotheses.

Denote generated hypotheses as $H = \{h_1, h_2, \dots, h_N\}$ and form a table like

$G \mid H$	h_1	h_2		h_N
g_1	0	1		0
:	•••	:	·	:
g_n	0	0		1
g_{n+1}	0	0		0
:	•••	:	0	:
$g_{(1+\delta)n}$	0	0		0
$g_{(1+\delta)n+1}$	1	0		0
:	•••	:	·	:
g_{2n}	0	1		1

Here 1 corresponds to inclusion of given hypothesis into a given example, i.e. the hypothesis predicts the example correctly (positively).

To reach the empirical risk $\eta = 0$ each of the first *n* rows must contain at least one 1.

Due to the uniform distribution and independence of hypotheses, the corresponding cells form an independent Bernoulli series with a probability of success $\frac{1}{2}$.

Lemma 4. If hypotheses number $N \ge (1+\sigma) \log_2 n$ for some $\sigma > 0$ then $\lim_{n\to\infty} \mathbb{P}[\eta = 0] = 1$. *Proof.*

$$1 \ge \lim_{n \to \infty} (1 - 2^{-N})^n = \lim_{n \to \infty} \left[(1 - 2^{-N})^{2^N} \right]^{n \cdot 2^{-N}} = \lim_{n \to \infty} \left[e^{-1} \right]^{n \cdot 2^{-N}} \ge \lim_{n \to \infty} e^{-1/n^{\sigma}} = 1.$$

To achieve $\theta = \delta$ fraction of erroneous predictions of test examples it needs to select $\delta \cdot n$ rows of the lower half (there are $\binom{n}{\delta \cdot n}$ ways to do this) containing zeroes only, the rest rows can contains ones somewhere. Table above corresponds to the situation with choice of rows $g_{n+1}, \ldots, g_{(1+\delta) \cdot n}$.

The task is to estimate $\mathbb{P} = {n \choose \delta \cdot n} \cdot (1 - 2^{-N})^{(2-\delta)n} \cdot (2^{-N})^{\delta \cdot n}$ when $n \to \infty$. Lemma 5 (Stirling's formula). $n! \approx \sqrt{2\pi n} \frac{n^n}{e^n}$ for $n \to \infty$. **Lemma 6** (Entropy inequality). $-\delta \cdot \ln \delta - (1-\delta) \cdot \ln(1-\delta) \le \ln 2$.

Theorem 1. $\lim_{n\to\infty} \mathbb{P} \leq \frac{1}{\sqrt{2\pi\delta(1-\delta)}} \exp\left\{-(1+\sigma)\cdot\delta\cdot n\cdot\ln n + \ln 2\cdot n - \frac{\ln n}{2}\right\}$ for $n\to\infty$ and $N \geq (1+\sigma)\log_2 n$.

Proof. The second factor of $\mathbb{P} = \binom{n}{\delta \cdot n} \cdot (1 - 2^{-N})^{(2-\delta)n} \cdot (2^{-N})^{\delta \cdot n}$ does not exceed 1. Stirling's formula and Entropy inequality imply

$$\mathbb{P} \leq \frac{\sqrt{2\pi \cdot n} \cdot n^{\delta \cdot n} \cdot n^{(1-\delta) \cdot n} \cdot e^{\delta \cdot n} \cdot e^{(1-\delta) \cdot n} \cdot (2^{-N})^{\delta \cdot n}}{e^{n} \cdot \sqrt{2\pi\delta \cdot n} \cdot \sqrt{2\pi(1-\delta) \cdot n} \cdot \delta^{\delta \cdot n} \cdot n^{\delta \cdot n} \cdot (1-\delta)^{(1-\delta) \cdot n} \cdot n^{(1-\delta) \cdot n}} \leq \frac{2^{-\delta \cdot n \cdot (1+\sigma) \cdot \log_2 n}}{\sqrt{2\pi\delta(1-\delta) \cdot n} \cdot e^{\delta \cdot n \cdot \ln \delta} \cdot e^{(1-\delta) \cdot n \cdot \ln(1-\delta)}} = \frac{e^{-\delta \cdot (1+\sigma) \cdot n \cdot \ln n} \cdot e^{n \cdot (-\delta \cdot \ln \delta - (1-\delta) \cdot \ln(1-\delta))}}{\sqrt{2\pi\delta(1-\delta)} \cdot \sqrt{n}} \leq \frac{e^{-\delta \cdot (1+\sigma) \cdot n \cdot \ln n} \cdot e^{\ln 2 \cdot n}}{\sqrt{2\pi\delta(1-\delta)} \cdot \sqrt{n}}$$

Conclusions

Main theorem of the paper provides a small upper bound on a probability of overfitting in FCA-based Machine Learning in the simplest case of Boolean algebra without counter-examples. It states that the probability of overfitting in fixed fraction of test examples tends to zero faster than exponent when the number of attributes goes to infinity. Interesting alternative for our research is the work of T.P. Makhalova and S.O. Kuznetsov [6], where classifiers form a lattice.

Acknowledgements.

The author would like to thank his colleagues at Dorodnicyn Computing Center of Russian Academy of Science and especially his PhD student L.A. Yakimova for collaboration and helpful discussions.

The author is grateful to anonymous reviewers for improving the style of presentation.

References

- [1] Ganter Bernard and Wille Rudolf. Formal Concept Analysis: Mathematical Foundations, Springer-Verlag, 1999
- [2] Diaconis Persi. *Group representations in probability and statistics*. IMS Lecture Notes Monograph Series Vol. 11.– Hayward (CA): Institute of Mathematical Statistics, 1988
- [3] Kuznetsov S.O. A Fast Algorithm for Computing all Intersections of Objects in a Finite Semi-Lattice. Automatic Documentation and Mathematical Linguistics. Vol. 27. – no. 5. – 1993. – pp. 11–21

- [4] Kuznetsov S.O. Complexity of Learning in Concept Lattices from Positive and Negative Examples. *Discrete Applied Mathematics*, no. 142(1-3). 2004. pp. 111–125
- [5] Kuznetsov S.O. Machine Learning and Formal Concept Analysis. Proc. 2nd International Conference on Formal Concept Analysis: Springer LNAI, Vol. 2961. – 2004. – pp. 287–312
- [6] Makhalova T.P. and Kuznetsov S.O. On Overfitting of Classifiers Making a Lattice. Proc. 14th International Conference on Formal Concept Analysis: Springer LNAI, Vol. 10308. – 2017. – pp. 184–197
- [7] Vapnik V.N. Statistical Learning Theory, Wiley-Interscience, 1998
- [8] Vinogradov D.V. A Markov Chain Approach to Random Generation of Formal Concepts. Proc. of Workshop Formal Concept Analysis Meets Information Retrieval (FCAIR 2013): CEUR Workshop Proceedings, Vol. 977. – 2013. – p. 127–133
- [9] Vinogradov D.V. Accidental Formal Concepts in the Presence of Counterexamples. Proc. of Workshop on Formal Concept Analysis for Knowledge Discovery (FCA4KD 2017): CEUR Workshop Proceedings, Vol. 1921. – 2017. – p. 104–112
- [10] Vinogradov D.V. FCA-based Approach to Machine Learning. Proc. of Workshop on Formal Concept Analysis for Artificial Intelligence (FCA4AI 2019): CEUR Workshop Proceedings, Vol. 2529. – 2019. – p. 57–64
- [11] Vorontsov K.V. and Ivahnenko A. Tight Combinatorial Generalization Bounds for Threshold Conjunction Rules. Proc. of 4th International Conference on Pattern Recognition and Machine Intelligence. – 2011. – p. 66–73