人工知能学会研究会資料
SIG-AGI-006-06

Solomonoffの万能推論・アルゴリズム的確率

Solomonoff's Universal induction, or Algorithmic probability

宮部賢志 [1*]

[1] 明治大学理工学部数学科
[1] Department of Mathematics, Meiji University

Abstract: We give an introduction of Solomonoff's universal induction, or algorithmic probability. The existence of universal prior (or computability) is the key of his result, which explains many aspects in artificial intelligence and philosophy of science. This introduction especially focuses on Solomonoff's view of probability.

人工知能を作る上では，人間がどのように学習をしているのかを観察する．科学哲学では，理論の選択基準や予測の意味を考察する．数学では，確率という概念をどう理解し定式化するかという問題が論じられた．その結果，学習や予測についても特定の設定の上では数学的に論じることができるようになった．本論文では，von Mises や Kolmogorov の確率論の形成について簡単に振り返りながら，Solomonoff の確率や予測の概念の哲学的・数学的意味について概説する．

1 小史

Hilbert の 23 の問題は，Hilbert により 1900 年の国際数学者会議とその後の著作において提案された当時未解決であった問題群である．その第 6 問題は「物理学は公理化できるか」である．物理学 (physics) とあるが，特に確率概念の公理化が念頭に置かれている．確率という概念はその意味の議論が昔から絶えない．それだけでなく，数学的な取り扱いについて混乱していた時代でもあった．

確率の公理化という問題に満足行く形で答えたのは，Kolmogorov (1903-1987) の『確率論の基礎概念』[6] であった．この確率論は測度論に基づいており，確率の性質だけを問題にしていたため，頻度説，主観説，傾向説など確率について様々な立場を持つ人にも受け入れられ，現代確率論の基礎を立脚することになった．

ところが，Kolmogorov 自身は頻度論者であり，自身の公理系に満足していなかったことが，以下の文章から分かる．1963 年の論文 [7] からの引用である．

> I have already expressed the view that the basis for the applicability of the results of the mathematical theory of probability to real 'random phenomena' must depend on some form of frequency concept of probability, the unavoidable nature of which has been established by von Mises in a spirited manner.

von Mises (1881-1973) は確率概念を頻度説を通じて定式化しようと試みた．その試みは 1919 年の論文 [21] に始まり，大部な本 [22] になっているが，その核心部分は薄い本 [23] としてまとめられている．

1940 年代ごろまでは Kolmogorov の体系と von Mises の理論のどちらを採用すべきかという議論も行われていたようであるが，最終的にはその簡潔さから Kolmogorov の体系が標準的な確率論とみなされている．しかし，von Mises の理論に学ぶことは多い．

von Mises の考えは，"First collective, then probability" の言葉でまとめられる．通常の測度論的確率論では，確率の定まった空間を設定し[1]，その実現値としての標本のランダム性を調べる．それに対し，collective は大雑把に言えばランダムな列のことであり，その極限頻度として確率が定まるのだというのが von Mises の考えである．通常の確率論とは方向が正反対であることに注意しよう．Ville [20] が指摘したように，collective はランダムな列の定義としては不十分であった．

ランダムな列の定義を満足の行く形で示したのは Martin-Löf [10] であった．その本質は統計的仮説検定の計算可能性を考えるというものである．更に Kolmogorov 複雑性という記述不可能性による特徴付けが得られたことから，文字列の情報量という考え方の有用性が理解されるようになった．今日，アルゴリズム的情報理論 (algorithmic information theory) と呼ばれており，Kolmogorov はその創始者である．ランダム性

*連絡先：明治大学理工学部数学科
〒 214-8571 神奈川県川崎市多摩区東三田 1-1-1
E-mail: research@kenshi.miyabe.name

[1] 伊藤清の『確率論の基礎』では「確率とは，ルベーグ測度である」と表現されている．

についてのその後の発展については，[9, 13, 1] などが主要文献である．

Solomonoff (1926-2009) は人工知能への応用を見据えて「確率論」を捉えなおそうとした．その考えは von Mises や Kolmogorov らの発展と見るのが理解しやすい．次節以降では Solomonoff の考えの以下の点に特に注目して解説する．

1. データから出発して，そのランダム性に着目して，確率を定める
2. 万能機械の存在から，(極限においては) 最適な予測の存在が導かれる．

この分野の主要文献としては [3] が挙げられる．また，哲学的な重要性としては [15, 17] などを見よ．

Solomonoff の考えは最小記述量の理論などへも影響を与えているが，いくつかの困難性から十分発展しているとは言い難い．今後の発展の可能性についても最後に議論する．

2 アルゴリズム的確率の考え方

2.1 設定と問題

まずは以下の設定で考えよう．考える空間は Cantor 空間 $\{0,1\}^{\mathbb{N}}$ である．ある元 $X \in \{0,1\}^{\mathbb{N}}$ を予測するしたい．正確には，その最初の n 桁である $X \upharpoonright n = X(0)X(1)\cdots X(n-1)$ が与えられた時に，$X(n) = 0, 1$ それぞれの賭け率はどうすべきか．

気持ちとしては $X(n) = 0, 1$ のそれぞれの確率を求めるという問題である．しかし，Solomonoff の考えに従い，確率という言葉は特別な意味で使いたいので，ここでは賭け率や測度など標準的でない表現を使う．

ベイズ予測の設定に従い，$\{0,1\}^{\mathbb{N}}$ 上に予測確率としての測度 M を定めれば，$\sigma \in \{0,1\}^*$ が与えられた時に，次が $i \in \{0,1\}$ である掛け率は，

$$M(i|\sigma) = \frac{M(\sigma i)}{M(\sigma)}$$

で与えればよいであろう．では，この M はどう選べばよいだろうか？

2.2 予測への計算可能性という制約

X の分布について何か仮定を置きたくなる．例えば，すべてのビットが独立同分布であるなどの条件を置けば，それまでの文字列の $0, 1$ の個数を数えることで，良い予測ができるだろう．しかし，私たちはそのような仮定をできる限り置きたくない．X の分布に制限を置くことは，数理モデルや理論に仮定を置くことに相当する．限られたモデルの中から良いパラメータを選ぶのではなく，モデルそのものを探す方法を考えたいからである．

M に対してある意味で計算可能 (正確には異なる．後に修正する．) という仮定を置く．M は 2 進有限列から実数への関数と見ることができ，その関数が計算可能なものの中から選ぶことにする．計算可能であるとは,「適当なプログラミング言語で表現できる」と思えば良い．この制約は妥当であろう．人工知能や予測の問題を考える上では，その予測は計算可能であって欲しい．この制約は理論構築上，技術的にも重要である．確率測度全体の濃度は非可算であるのに対し，計算可能な測度全体の濃度は可算なので，考える対象がずっと少なくなる．それでいて，表現力は強く，かなり複雑な予測を表現することができる．

予測確率を直接議論していることにも注意してほしい．一般的には「データが与えられて，それを説明するモデルを考えて，そのモデルの予測を我々の予測とする」という考え方をすることが多い．その場合，モデルが計算可能であるべきかとか，確率モデルを採用して良いのか，といった問題を考える必要がある．モデルの確信度合いにより，予測にもゆらぎが生じる．どんなモデル群を考えていたとしても，データから予測を与えるのであれば，それは関数と見ることができる．モデルフリーの予測と見ることもできるが，プログラムとモデルを同一視していると見ることもできる．この点については後に議論する．

2.3 万能機械と万能予測の存在

部分関数 $f :\subseteq \mathbb{N} \to \mathbb{N}$ が計算可能であるとはどういうことか．Hilbert の第 10 問題は「ディオファントス方程式 (整数係数の多項式の方程式) に対して，整数解が存在するかどうかを判定するアルゴリズムを与えよ」という問題であった．当時はそのようなアルゴリズムはきっと存在するだろうと固く信じられていた．そのようなアルゴリズムが存在しないことを示すには，計算可能関数の数学的定義が必要である．この計算可能関数の定義に決着を与えたのは，1936 年の Turing の論文 [19] であった．今日では，計算の理論 (the theory of computation) として発展している．この分野の文献としては，日本語にも翻訳されている Sipser [16] などがある．Church-Turing のテーゼより，関数 $f :\subseteq \mathbb{N} \to \mathbb{N}$ が計算可能であるとは，f が Turing 機械で計算可能であることと定義する．とりあえずは適当なプログラミング言語で書くことができると理解すれば良い．

計算の理論での重要な結果として，万能機械 (universal machine) の存在が挙げられる．例えば，電卓は足

し算や掛け算など特定の計算はできるが，どんな計算でもできるというわけではない．現在の計算機はプログラムさえダウンロードすれば，時間やメモリの制限を無視すれば，どんな計算でもできる．そのような機械が存在しうるということは Turing の論文の結果であり，それを実現したものが現在の計算機である．万能機械はどんな計算も模倣できる機械であるから，万能機械で計算できない関数はどんな機械でも計算できない．

万能機械が存在するのと同様に，万能な予測 M が存在する．技術的な理由により，考える予測確率の対象を少し広げる．関数 $m : \{0,1\}^* \to [0,1]$ が**半測度** (semimeasure) とは，$m(\lambda) \leq 1$ かつすべての $\sigma \in \{0,1\}^*$ に対して，$m(\sigma) \geq m(\sigma 0)+m(\sigma 1)$ であることをいう．λ は空列 (長さ 0 の文字列) を表す．2 つの不等号が等号に置き換われば測度となる．実数 $x \in \mathbb{R}$ が**下側半計算可能** (lower semicomputable) であるとは，計算可能な単調増加の有理数列 $\{a_n\}$ が存在して，$x = \lim_n a_n$ となることをいう．すなわち，下側から計算可能に近似できる数のことである．この定義は $f : \{0,1\}^* \to [0,1]$ の形の関数にも自然に拡張できる．

定理 2.1 (万能予測の存在)．以下の性質を満たす下側半計算可能な半測度 $M : \{0,1\}^* \to [0,1]$ が存在する：すべての下側半計算可能な半測度 m に対して，ある定数 $c \in \mathbb{N}$ が存在して，すべての $\sigma \in \{0,1\}^*$ に対し，

$$cM(\sigma) \geq m(\sigma)$$

を満たす．

M は $X \in \{0,1\}^{\mathbb{N}}$ を予測している．$M(\sigma)$ は X が σ から始まると思う信念の度合いを表している．その $M(\sigma)$ がどんな m の予測 $m(\sigma)$ と比較しても，$1/c$ 倍よりは悪くならないということを意味している．この万能性の主張は c が含まれているためあまりにも弱いと感じるかもしれない．私たちが興味のあるのは，M そのものではなく，その比 $\dfrac{M(\sigma i)}{M(\sigma)}$ であるため，極限においてはこの c は大きな問題にならない．

M は計算可能に近似できるが，計算可能ではない．上記の定理において，下側半計算可能な半測度を計算可能な測度に置き換えると成り立たなくなる．

2.4 アルゴリズム的確率

上記の万能予測 M を 1 つ固定し，その比 $\dfrac{M(\sigma i)}{M(\sigma)}$ を**アルゴリズム的確率** (algorithmic probability) と呼ぶ．確率をこの形のものに限定するのは実に使い勝手が悪い．そのため，アルゴリズム的確率という言葉を使うのは Solomonoff 自身くらいで，他の研究者は万能推論 (universal induction) と呼ぶことが多い．

3 複雑性による表現

3.1 良い理論とは

アルゴリズム的確率はモデル選択の観点から解釈することができることを説明する．

ある現象を説明する複数の理論に対して，一見正反対の 2 つの考え方がある．

1 つは Epikuros による「それらが現象と矛盾しない限りすべて保持せよ」というもの．理論同士は矛盾しているかもしれないが，現象の観察が進むにつれて，正しくないものはやがて分かるだろうという考え方である．

もう 1 つは Occam の剃刀といわれる「最も単純なものを採用せよ」というもの．理論はより単純である方が良い予測をすることが多い．この考え方は情報量規範という形で，現在も広く使われている．

3.2 モデルの足し合わせ

アルゴリズム的確率ではモデルとプログラムを同一視する．そして，現在までに与えられた情報と整合性のあるプログラムすべてを保持しつつ，そのプログラムの長さが短いものほど可能性が高いと考える．すべての予測は実数値の関数なので，それらを足し合わせたものを最終的な予測とする．

以上の考え方を数式にすると，以下のようになる：

$$M(\sigma) = \sum \{2^{-|p|} \ : \ U(p) = \sigma *\}$$

ここで U は単調万能機械 (monotone universal machine) であり，$p \in \{0,1\}^*$ はプログラムを想定している．$U(p) = \sigma *$ は $U(p) \in \{0,1\}^{\leq \mathbb{N}}$ が σ から始まることを表す．$|p|$ は文字列 p の長さである．よって，長さが短いほど可能性が高いとして，足し合わせていることになる．予測を関数として見ることは，このような足し合わせが容易になるという利点もある．

実はこの M は下側半計算可能な半測度となり，前節で述べた意味で万能な予測であることが証明できる．アルゴリズム的確率の立場としては，M が万能であるということだけが重要で，モデルという考え方を経由しているかどうかは重要ではない．しかし上記のような意味で，モデルの選択をしているという見方をすることもできる．

4 アルゴリズム的確率の性質

上記の意味で万能性を持つ予測 M から導かれるアルゴリズム的確率は，本当に予測として良い性質を持っているのだろうか？以下ではアルゴリズム的確率の持つ性質を持っているのかを見ていく．

4.1 計算可能な列の予測

最も単純な例から考えよう．あるゲーム (例えば将棋など) において，n 連勝している人がいたとして，次のゲームで勝つ確率を計算したい．単純に頻度を考えて「それまですべて勝ったのだから，次に勝つ確率は 100%」とするのは，不自然である．n が大きければ大きいほど，確率は 1 に近づくはずである．ではその収束の速さはどの程度だろうか?この問題は Raven paradox に関連して昔から論じられてきた問題である．

勝ちを 1 で表し，負けを 0 で表す．この人が無限回ゲームを行った時の勝敗の結果が $X \in \{0,1\}^{\mathbb{N}}$ である．求める確率は $M(1|1^n) = \dfrac{M(1^{n+1})}{M(1^n)}$ である．

Hutter [4] は，一般に計算可能な列 $X \in \{0,1\}^{\mathbb{N}}$ に対して，ある $C \in \mathbb{N}$ が存在して，

$$2^{-K(n)} \leq 1 - M(X_n|X_{<n}) \leq C \cdot 2^{-K(n)}$$

となることを示している．ここで，K は Kolmogorov 複雑性であり，万能機械 U に対して，

$$K(\sigma) = \min\{|\tau| \ : \ U(\tau) = \sigma\}$$

で定義される．

上記の例の場合，$\lim_n K(n) = \infty$ なので，

$$\lim_n M(1|1^n) = 1$$

であり，確率は 1 に近づく．ほとんどの複雑な n については，$K(n) \geq \log n + \log\log n$（ただし，log の底は 2）なので，$1 - M(1|1^n)$ は $1/n$ よりも速く 0 に収束する．しかし，単純な n については，誤差は $1/n$ とくらべてかなり大きくなる．このような振動は計算可能性を考えなければあまり見られない現象であろう．

$X = 1^{\mathbb{N}}$ の場合には，人間でもその規則性にやがて気がつくだろう．一方，例えば X が π の小数部分の 2 進展開などの場合，人間ではその規則性に気がつくのはかなり困難である．しかし，上記の M は上記の意味で正しく予測できる．すなわち，M は全く事前情報なしでも，計算可能な列の規則を見つけることができるのである．

4.2 ランダム列の予測

次に，X がランダムである場合を考えよう．例えば，偶数番目は 1 が 1/3 の確率で，奇数番目は 1 が 2/3 の確率で出るモデルを考える．人間がその規則を見つけるのはかなり難しい．では，その標本に沿って予測を行った時に，その M は正しくその確率に収束するだろうか？

これを数学的に表現するために，アルゴリズム的確率と異なる確率の概念を考える．$\{0,1\}^{\mathbb{N}}$ 上の計算可能な確率測度 m を考える．X をその上の標本とする．この時，m の意味での確率 1 で

$$\lim_n |M(X_n|X_{<n}) - m(X_n|X_{<n})| = 0$$

を満たす [18]．すなわち，やはり M は事前情報なしで，m の規則性を見つけることができるのである．

この m を確率モデルと思うならば，確率 1 で正しい確率に収束すると思うことができる．ここで出て来る確率は異なる意味であることに注意しなければならない．アルゴリズム的確率が $M(X_n|X_{<n})$ であるならば，m は一体何なのだろうか？この文脈では m を確率の言葉で表現するのは適切ではない．

アルゴリズム的ランダムネスの理論では，列そのものがランダムであることの定義を与えることができる．よって，列というデータから出発して，その収束先としての確率を考えることができるのである．ランダムネスの理論における中心的な概念は，Martin-Löf ランダムネスである．上記の結果はすべての ML ランダムな列で収束すると間違って信じられてきた．最近になって，Hutter ら [5, 8] により ML ランダムな列では不十分であることが示されている．もちろん，より強いランダムの概念であれば収束する．

5 課題

Solomonoff のアルゴリズム的確率は，計算可能性の概念を利用することにより，自然な予測限界を明らかにしている．その哲学的意味も大きく，通常の確率とは違った見方を与えている．今後の発展の可能性として考えられることを挙げよう．

最も大きな問題は中心的な役割を果たす M が計算可能ではないことであろう．そのため，そのまま応用できるわけではない．この事実は計算論全般で同じく問題になっている．例えば，正規化圧縮距離のように，計算不可能な複雑性を適当な圧縮形式のサイズで置き換えるなどの大胆な近似が必要であろう．

本稿では 2 進の文字列において考察した．この設定は自然に有限種類のアルファベットに拡張できる．実数などのより一般的な状況に拡張して，他の手法との比較を行う必要があるだろう．筆者の以前の論文 [11] はこの方向性を目指したものである．

M が正しい確率に収束するような列の集合の特徴付けはまだまだ不十分である．ランダムネスの理論で Cantor 空間上の一様測度においては，様々なランダムの概念が研究されている．特に最近研究されている密度ランダムネスの研究 [12] は，この M の収束する列と深い関係があるように思われる．

M の計算可能性と予測できる確率 (正しい確率に収束するような確率) との間には深い関係がある．この関係の研究には Levy の 0-1 法則の計算可能性について考察する必要がある．関連する最近の研究として，Lebesgue の微分定理の計算可能性 [14] や，Radon-Nikodym の定理の計算可能性 [2] の研究がある．また，計算量の観点から対応する階層を研究することも，本質的に重要な方向性であると思われる．

アルゴリズム的確率という観点からすれば，測度 m が最初に与えられるのはやはり不自然である．データ X に対して M による予測が収束する確率のより自然な解釈がないだろうか．

謝辞

本稿は科研費 (26870143) の助成を受けた研究に基づいたものである．

参考文献

[1] R. Downey and D. R. Hirschfeldt. *Algorithmic Randomness and Complexity*. Springer, Berlin, 2010.

[2] M. Hoyrup, C. Rojas, and K. Weihrauch. Computability of the Radon-Nikodym derivative. *Computability*, 1:3–13, 2012.

[3] M. Hutter. *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer, 2005.

[4] M. Hutter. On Universal Prediction and Bayesian Confirmation. *Theoretical Computer Science*, 384:33–48, 2007.

[5] M. Hutter and A. Muchnik. On semimeasures predicting Martin-Löf random sequences. *Theoretical Computer Science*, 382:247–261, 2007.

[6] A. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, 1933.

[7] A. N. Kolmogorov. On tables of random numbers. *Sankhyā: The Indian Journal of Statistics, Series A*, 25(4):369–376, 1963.

[8] T. Lattimore and M. Hutter. On Martin-Löf (non-)convergence of Solomonoff's universal mixture. *Theoretical Computer Science*, 588:2–15, 2015.

[9] M. Li and P. Vitányi. *An introduction to Kolmogorov complexity and its applications*. Graduate Texts in Computer Science. Springer-Verlag, New York, third edition, 2009.

[10] P. Martin-Löf. The Definition of Random Sequences. *Information and Control*, 9(6):602–619, 1966.

[11] K. Miyabe. An optimal superfarthingale and its convergence over a computable topological space. *Lecture Notes in Artificial Intelligence*, 7070:273–284, 2013.

[12] K. Miyabe, A. Nies, and J. Zhang. Using almost-everywhere theorems from analysis to study randomness. *The Bulletin of Symbolic Logic*, 22(3):305–331, 2016.

[13] A. Nies. *Computability and Randomness*. Oxford University Press, USA, 2009.

[14] N. Pathak, C. Rojas, and S. G. Simpson. Schnorr randomness and the Lebesgue Differentiation Theorem. *Proceedings of the American Mathematical Society*, 142:335–349, 2014.

[15] S. Rathmanner and M. Hutter. A Philosophical Treatise of Universal Induction. *Entoropy*, 13:1076–1136, 2011.

[16] M. Sipser. *Introduction to the Theory of Computation*. Course Technology Ptr, second edition edition, 2012.

[17] R. Solomonoff. Algorithmic probability: Theory and applications. *Information Theory and Statistical Learning*, pages 1–23, 2009.

[18] R. J. Solomonoff. Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Transaction on Information Theory*, IT-24:422–432, 1978.

[19] A. M. Turing. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 420:230–265, 1936.

[20] J. Ville. Étude critique de la notion de collectif. *Gauthier-Villars*, 1939.

[21] R. von Mises. Grundlagen der Wahrscheinlichkeistrechnung. *Mathematische Zeitschrift*, 5:52–99, 1919.

[22] R. von Mises. *Mathematical theory of probability and statistics.* Academic Press Inc, 1964.

[23] R. von Mises. *Probability, statistics, and truth.* Dover Pubns, 1981.