



入門講座

インフォマティクス入門ー7

ものづくり分野における能動学習

Active Learning for Manufacturing

竹内一郎

Ichiro Takeuchi

名古屋工業大学
工学研究科 教授

理化学研究所 革新知能統合研究センター
データ駆動型生物医科学チーム
チームリーダー

1 はじめに

さまざまな科学技術分野で機械学習を用いたデータ駆動型アプローチが注目を集めている。機械学習には教師なし学習や強化学習など様々な問題設定があるが、本稿で紹介する能動学習 (Active Learning) は最も代表的な問題設定である教師あり学習の枠組の一つである。教師あり学習とは、入力データと出力データのペア (訓練事例と呼ばれる) が与えられた状況で、入出力関係を近似する関数をみつける問題である。科学技術分野の様々な問題が教師あり学習として定式化できる。例えば、実験科学において、実験条件を入力、実験結果を出力と考えると、これまでに行った実験から実験対象に関する知識を得るプロセスは教師あり学習と解釈することができる。教師あり学習では、より多くの訓練事例が利用可能であればあるほど、より複雑な関数を高精度に推定できる。例えば、さまざまな分野で活用が進む深層学習モデルは複雑な入出力関係を高精度に近似できることが知られているが、その学習には、大量の訓練事例が必要となる。

多くの実験科学では、実験に一定のコストがかかるため、膨大な訓練事例 (実験条件と実験結果のペア) を得るのは難しい。そのため、できるだけ少ない訓練事例で、できるだけ高精度に入出力関係を近似できるような工夫が必要となる。そのような工夫の一つとして、訓練事例を能動的に選択するアプローチが考えられる。これは、実験科学の例を用いれば、多数の実験条件 (入力) の候補のなかから、実際に実験を行う (関数を評価して出力を得る) ものを能動的に選択することを意味する。このように、訓練事例を得るコストが大きい場合、訓練事例を能動的に選択しながら機械学習を行うアプローチは能動学習と呼ばれている^{1,2)}。実験条件を系統的に決める一連の方法は実験計画法と呼ばれており、能動学習

は広い意味での実験計画の一種と解釈することができる。伝統的な実験計画では一連の実験を開始する前にすべての実験条件を決めてしまう場合が多いが、能動学習では実験対象のモデルを逐次的に更新しながら適応的に実験条件の選択が行われる。特に、材料科学分野では、材料情報学 (Material Informatics) と呼ばれる分野が注目を集めており、能動学習を用いた効率的な材料探索などの研究が行われている (例えば、3,4))。

2 問題設定

ここで、本稿で考える問題設定を整理しておこう。読者が直観的に理解しやすいように、仮想的な実験計画問題を考えることにしよう。可能な実験条件が N 通りあり、それらを、 $x_1, x_2, \dots, x_N$ と表すことにする。実験条件を x_i とした実験を実際に行って実験結果 y_i を得るプロセスを $y_i = f^*(x_i)$ と表すことにする。ここで、 $f^*: x_i \mapsto y_i$ は実験プロセス全体を表す関数であるので、 f^* がどのような関数であるのかは (実験を通して部分的にしか) わからないとし、 $f^*(x)$ を計算するコスト (すなわち、実験条件 x で実験を行うコスト) は大きいものとする。ここでの目的は、すべての N 通りの実験を行わず、できるだけ少ない実験回数で研究対象である実験プロセス f^* に関する知識を得ることである。この目的を実現するため、 N 通りの実験条件のうち、どの実験条件をどの順番で試すべきかを決めることを実験計画という。

以上の実験計画問題を能動学習の言葉で表現すると以下のようなになる。実験条件 $x \in X$ は入力変数、実験結果 $y \in Y$ は出力変数と呼ばれ、 x から y を得るための関数 $f^*: X \rightarrow Y$ が学習の対象となるブラックボックス関数である。ここで、 X と Y は、それぞれ、入力 x と出力 y のとりうる値の集合である。入

力集合 X が有限集合の場合 (すなわち、可能な実験条件がリストアップされている場合) をプール型能動学習と呼び、 X が連続集合の場合 (すなわち、可能な実験条件の範囲が与えられている場合) をクエリ型能動学習と呼ぶ。本稿では、 N 通りの実験条件を考えているため、入力集合が $X = \{x_1, \dots, x_N\}$ と表されるようなプール型能動学習を考えることになる。出力集合 Y が連続集合の場合を回帰問題、離散集合の場合を分類問題と呼ぶが、本稿では、出力空間が実数空間 $Y = \mathbb{R}$ である場合、すなわち、実験結果が数値として表される場合に限定して考察を行う。

以下、能動学習の基本的な流れを説明しよう。能動学習のステップを添字 $t = 1, 2, \dots$ を使って表すことにする。ステップ t において、すでに関数評価が行われた入出力事例、すなわち、すでに実験が行われた実験条件をラベルあり事例 (labeled instance) と呼び、まだ関数評価が行われていない入出力事例をラベルなし事例 (unlabeled instance) と呼ぶ*。ステップ t におけるラベルあり事例の集合を L^t 、ラベルなし事例の集合を U^t と表すことにする。以上の表記を用いると、能動学習の基本的な流れは図1のアルゴリズムのようになる。能動学習アルゴリズムを特徴づけるのはアルゴリズムの5行目における関数 $\alpha: X \rightarrow \mathbb{R}$ である。この関数 $\alpha(x)$ は獲得関数 (Acquisition Function) と呼ばれ、入力 x においてブラックボックス関数 f^* を評価して出力 $f^*(x)$ を得ることがどれほど有益であるかを定量化するものである。したがって、どのようなタスクを解くかによって、どのような獲得関数を使うべきかが異なる。本稿では、代表的な能動学習のタスクとして、関数推定 (Function Approximation)、最適化 (Optimization)、レベル集合推定 (Level Set Estimation) の3つを考え、それぞれのタスクにおいてどのような獲得関数を用いるべきかを説明する。

Input: 可能な実験条件: $\{x_1, \dots, x_N\}$
 1: 機械学習モデル f を初期化
 2: $t \leftarrow 0, L^t \leftarrow \emptyset, U^t \leftarrow \{1, \dots, N\}$
 3: **while** 停止条件を満たす **do**
 4: **for** $i \in U^t$ **do**
 5: 獲得関数 $\alpha(x_i)$ を計算
 6: **end for**
 7: $i^t \leftarrow \operatorname{argmax}_{i \in U^t} \alpha(x_i)$
 8: $y_{i^t} = f^*(x_{i^t})$
 9: $t \leftarrow t + 1$
 10: $L^t \leftarrow L^{t-1} \cup \{i^t\}, U^t \leftarrow U^{t-1} \setminus \{i^t\}$
 11: $\{(x_i, y_i)\}_{i \in L^t}$ を訓練データとして機械学習モデル f を再学習
 12: **end while**
Output: 研究対象 f^* に関する所望の知識

図1 能動学習の基本的な流れ

3 ガウス過程モデル

能動学習においては、ブラックボックス関数 f^* の代わりに数理モデル f を導入する。適切な獲得関数を設計するにはモデルの不確実性を定量化できることが望ましいため、ベイズ統計の考え方に基づくモデルが有益である。本稿では、ガウス過程モデル (Gaussian Process Model: GP Model) と呼ばれるベイズ統計モデルを例として用いることにする。ガウス過程モデルは、関数に対する正規分布と解釈でき、平均関数 $\mu(x)$ と共分散関数 $k(x, x')$ によって特徴づけられる。平均関数 $\mu(x)$ は f の平均的なふるまいを表し、共分散関数 $k(x, x')$ はカーネル関数とも呼ばれ、2つの入力 x と x' の類似度を表している。入力 x が数値ベクトルとして表現されている場合には、例えば、RBFカーネル

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\ell^2}\right)$$

などが用いられる。ここで、 ℓ は関数の滑らかさに関するハイパーパラメータである。一方、入力 x が数値表現できない場合 (例えば、文字列やグラフなどの構造データ) であっても、適切にカーネル関数を決めることができれば、ガウス過程モデルを利用することができる。

ガウス過程モデルでは、ベイズ統計の考え方に基づくモデルであり、事前知識を導入することができる。ラベルあり集合 (すでに実施した実験条件と実験結果) が与えられたもと、それぞれの入力 x に対する出力の予測分布は

$$f(x) \mid \{(x_i, y_i)\}_{i \in L} \sim \mathcal{N}(\mu_L(x), \sigma_L^2(x))$$

と正規分布として与えられる。ここで、平均 $\mu_L(x)$ と分散 $\sigma_L^2(x)$ はカーネル関数値やラベルあり集合の出力を用いて解析的に表現することができる。

図2はガウス過程モデルの例を示している。横軸と縦軸は、それぞれ、入力と出力を示しており、入力が近いほど共分散 (カーネル関数) が大きいことを想定している。 $N=25$ 個の四角は実験条件と実験結果のペアの集合 $\{(x_i, y_i)\}_{i=1}^N$ を表している。これらのうち、グレーの四角は実際に実験を行ったラベルあり事例を表し、白い四角はまだ実験を行っていないラベルなし事例を表している。ラベルあり事例に基づいて推定されたガウス過程モデルの予測平均と不確実性が、それぞれ、青色の曲線、青色の領域で表されている。ガウス過程モデルは、特定の入力に対する予測分布が正規分布で表されるようなモデルとなっている。図2には、3つの入力に対する予測分布が赤い曲線で示されている。

* ここで、ラベルとは関数の出力 y_i を意味している。これは、クラス分類問題において、出力 y_i がラベルと呼ばれることが多いためである

4 関数近似のための能動学習

ブラックボックス関数全体に関する知識を得たい場合は関数近似の問題と解釈することができる。関数近似とは、すべての入力におけるブラックボックス関数の値をできるだけ精度よく知ることである。関数近似問題は以下のように定式化される：

$$\hat{f} = \underset{f}{\operatorname{argmin}} \sum_{i=1}^n (f^*(x_i) - f(x_i))^2 \quad (1)$$

ここで、 f^* はブラックボックス関数である。プール型能動学習では、すべての実験を行えば、当然ながら、すべての関数値を知ることができる。関数近似のための能動学習の目的はできるかぎり少ない試行回数で、式 (1) の目的関数を小さくすることである。

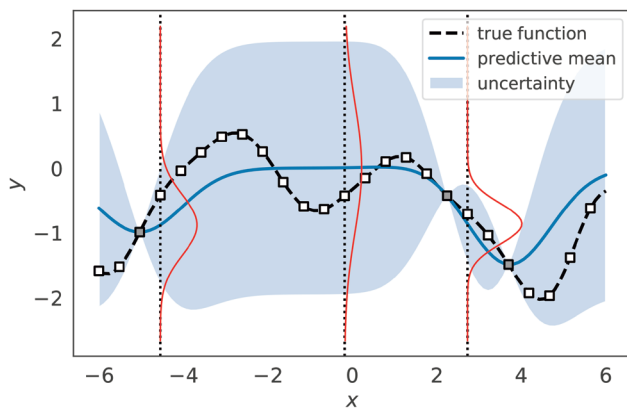


図2 ガウス過程モデルの例

ガウス過程モデルにおいては、各入力に対する不確実性が分散関数によって定量化されているので、関数近似のための能動学習における獲得関数では、不確実性サンプリング (Uncertainty Sampling) と呼ばれる獲得関数

$$\alpha_{\text{US}}(x) = \sigma_L(x)$$

が用いられることが多い。ラベルなし集合に含まれるすべての入力におけるガウス過程モデルの不確実性を計算し、それが最大となるような入力を選ぶことになる。

図3は不確実性サンプリングの獲得関数の概念と簡単なデモを表したものである。不確実性サンプリングでは、青い領域で示されているガウス過程モデルの不確実性をそのまま獲得関数として用いる。図のデモからわかるように、不確実性サンプリングでは、できるだけ均等に入力が選ばれる。これは、ガウス過程モデルの分散関数がカーネル関数によって定量化された共分散によって特徴づけられるため、すでに観測されたラベルあり集合の点からより遠い点の分散が大きくなるためである。また、ガウス過程モデルの分散関数 $\sigma(x)$ はラベルあり集合の観測値 y_i , $i \in L$ に依存していないことに留意されたい。これは、関数推定においては、必ずしも逐次的に入力の選択を行う必要がないことを示唆している。関数評価を行う回数 (何回実験ができるか) が決まっていれば、入力の集合をあらかじめ決めることができる。実際、不確実性サンプリングの考え方は、機械学習分野で新たに考案されたものでなく、伝統的な実験計画法分野で古くから確立されているものである。

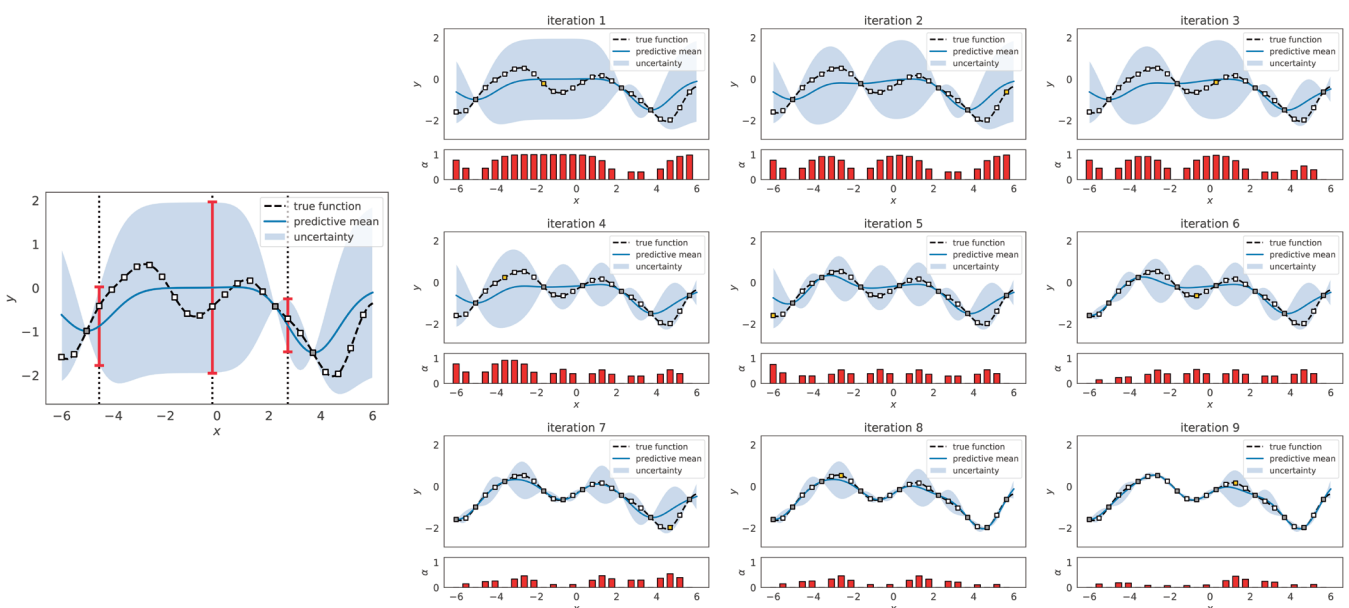


図3 (左) 不確実性サンプリングの獲得関数の概念図。不確実性サンプリングでは、不確実性の大きな点を選択される。
(右) 不確実性サンプリングにおけるガウス過程モデルと獲得関数の推移の例

5 最適化のための能動学習

必ずしもブラックボックス関数のすべてを知る必要がなく、なんらかの指標が最適となるような入力（実験条件）をみつきたい状況では、最適化のための能動学習が有益である。ガウス過程モデルに基づく能動学習において最適化を行う枠組はベイズ最適化 (Bayesian Optimization) と呼ばれ、活発に研究が行われている^{5,6)}。例えば、材料科学分野における材料探索問題では、もっとも高い物性値を持つ材料をできるだけ少ない実験回数で見つけ出したいというニーズがある。また、機械学習分野では、モデルの最適なハイパーパラメータをできるだけ少ない回数の学習で見つけ出したいというニーズがあり、ベイズ最適化はその基本ツールとなっている。

最適化のための能動学習は、入力候補点のなかから、最もブラックボックス関数値が大きくなるものを見つけることであるので、

$$\hat{x} = \operatorname{argmax}_{x \in X} f(x)$$

をできるかぎり少ない試行回数で求めることが目的となる。最適化のための能動学習では、これまでに見つかった最良の値 (Current Best : CB)

$$y_{CB}^t = \max_{i \in L^t} f^*(x_i)$$

よりもよい値をとりうる入力を探すことになる。最適化のための能動学習には様々な指標が提案されているが代表的なものは以下の2つである。1つ目は改善確率 (Probability of Improvement (PI)) と呼ばれるもので、 y_{CB}^t よりもよい値を

とる確率を指標とする。入力 x に対する予測分布が $f(x) | \{(x_i, y_i)\}_{i \in L^t} \sim \mathcal{N}(\mu_L(x), \sigma_L^2(x))$ と表されることを利用して

$$\alpha_{PI}(x) = \Pr(f(x) > y_{CB}^t | \{(x_i, y_i)\}_{i \in L^t})$$

と定式化される。

PIに基づく基準では、 y_{CB} より少しでもよい出力をとる確率が高い場合にも獲得関数の値が大きくなるが、問題によっては、新たな実験による改良の程度が重要なこともある。そのような場合、改善期待値 (Expected Improvement : EI) と呼ばれる指標が使われる。EI指標は

$$\alpha_{EI}(x) = \mathbb{E}[\max\{0, f(x) - y_{CB}^t\}]$$

と定式化される。

ベイズ最適化の獲得関数について議論される際、探索 (exploration) と搾取 (exploitation) のトレードオフを考慮する必要がある。探索 (exploration) とは不確実性の高いところを重視する考え方で、搾取 (exploitation) とは関数値が大ききところを重視する考え方である。もしも、ブラックボックス関数モデルの平均関数のみしか使わなければ、平均関数値が最大となる入力を選ぶ選択肢しかないが (搾取)、分散関数を使えば不確実性が大きいために改善の可能性が残されている点を選ぶこと (探索) も可能である。

図4は最適化のための獲得関数の概念と簡単なデモを表したものである。最適化問題においては、現時点での最良解 y_{CB}^t より大きな値をとる確率やその期待値に基づいて入力を選択されるため、左の図の赤く塗られた部分の確率や期待値が獲得関数となる。図のデモからわかるように、最適化のための

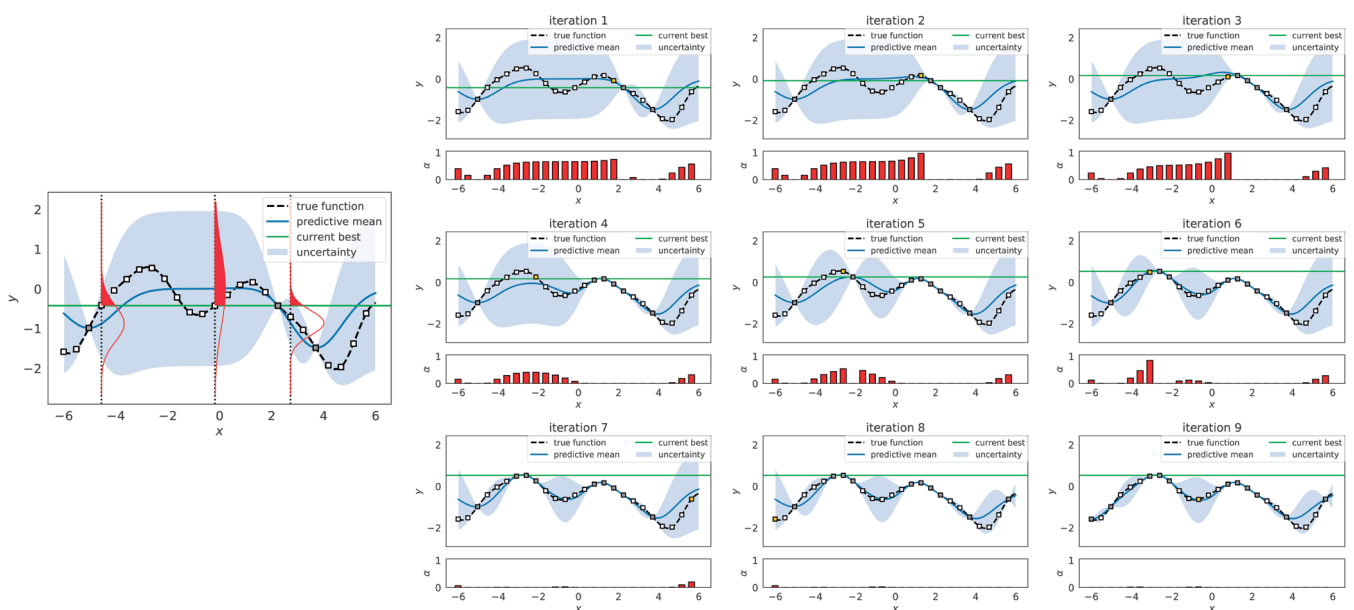


図4 (左) ベイズ最適化のための獲得関数(Probability of Improvement:PI) の概念図。この獲得関数では、これまでの最良のものよりもよい値が得られる確率が最大の入力を選択される。(右) ベイズ最適化(PI) におけるガウス過程モデルと獲得関数の推移の例

能動学習では、図3の関数近似問題と異なり、値が低いところでは不確実性があっても選択されず、値が高いところを重点的に探索していることがわかる。

6 レベル集合推定のための能動学習

様々な分野において、特定の性能を満たすような入力条件を同定したい問題がある。例えば、製品の性能が一定値以上となるような使用条件をみつけることは工業製品の品質管理の基本である。このような問題はレベル集合推定 (Level Set Estimation : LSE) と呼ばれ、ブラックボックス関数値 $f^*(x)$ がある閾値 θ よりも大きな入力の集合

$$\{x \in X | f^*(x) \geq \theta\}$$

をできるかぎり少ない試行回数で推定することが目的となる^{7,8)}。ここで、 θ はあらかじめ指定される閾値であり、 $f(x) \geq \theta$ となるような入力の集合を求めることが目的となる。

LSEのための能動学習法には複数の方法が提案されているが、本稿では、最も基本的なアプローチである Straddle と呼ばれる方法を紹介する。レベルセット推定の目的は、各入力における出力が閾値 θ より大きいか小さいかを推定することであるので、その不確実性は

$$\alpha_{\text{LSE}}(x) = \min \left\{ \max \left\{ 0, (\mu_L(x) + \beta \sigma_L^2(x)) - \theta \right\}, \max \left\{ 0, \theta - (\mu_L(x) - \beta \sigma_L^2(x)) \right\} \right\}$$

と定式化でき、これをそのまま獲得関数として利用できる。

ここで、 $\beta > 0$ はハイパーパラメータであり、例えば、 $\beta = 3$ とすると、出力の不確実性を $\pm 3\sigma$ の範囲で考えることに相当する。この指標は、閾値を下回ると思われる場合に関数値が上回る可能性、閾値を上回ると思われる場合に関数値が上回る可能性を表しており、すべての入力においてこの指標が 0 となった段階で、(指定された信頼度で) すべての入力が閾値を下回るか上回るかを確定できたことになる。

図5はレベルセット推定のための獲得関数の概念と簡単なデモを表したものである。図のデモからわかるように、レベルセット推定のための能動学習では、閾値より明らかに小さな値をとる点や大きな値をとる点は、たとえ不確実性が大きくても選ばれず、閾値付近の値をとり上回るか下回るか不明な点を重点的に選択していることがわかる。

7 まとめ

様々な分野で機械学習を活用したデータ駆動型アプローチが有望視されているが、ものづくりなど訓練事例の取得にコストがかかる分野では、能動学習が有効に利用できる。能動学習は訓練事例を能動的に選択しながら機械学習を行う枠組であるが、入力を実験条件、出力を実験結果と考えると、実験計画法的一种と解釈することもできる。能動学習のアルゴリズムは、次にどの入力を選ぶべきかを決定するための獲得関数によって特徴づけられるため、タスクに応じた獲得関数を設計する必要がある。獲得関数を設計する際には、ブラックボックス関数に関する不確実性を知ることが重要であるた

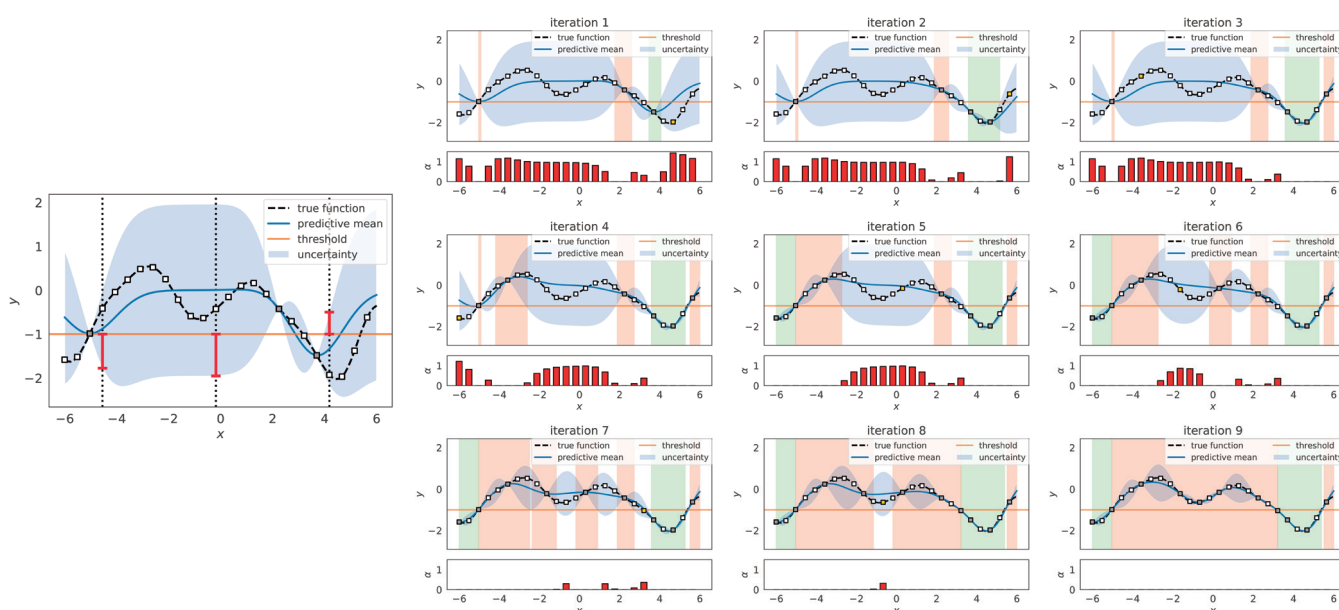


図5 (左) レベルセット推定のための獲得関数の概念図。レベルセット推定では、閾値を上回るか下回るかの不確実性が最大の入力を選択される。
(右) レベルセット推定におけるガウス過程モデルと獲得関数の推移の例。薄い赤色と緑色で示された領域は閾値を上回る可能性と下回る可能性が極めて高いと判定された領域を示している

め、ガウス過程モデルなどのベイズ統計の考え方に基づくモデルが有効利用できる。本稿では、関数近似、最適化、レベル集合推定の3つのタスクを考察し、それぞれのタスクに応じた獲得関数を使うことによって異なる入力を選ばれることを確認した。能動学習を活用することで、これまでは専門家の知識や経験に基づいて行ってきたことを、系統的かつ効率的に行えることが期待される。

参考文献

- 1) B.Settles : Active Learning Literature survey, Technical Report, University of Wisconsin-Madison Department of Computer Sciences, (2009)
- 2) Y.Inatsu, M.Karasuyama, K.Inoue, H.Kandori and I.Takeuchi : Neural Computation, 32 (2020) 10, 1998.
- 3) K.Kanamori, K.Toyoura, J.Honda, K.Hattori, A.Seko, M.Karasuyama, K.Shitara, M.Shiga, A.Kuwabara and I.Takeuchi : Phys. Rev. B, 97 (2018) 12, 125124.
- 4) R.Jalem, K.Kanamori, I.Takeuchi, M.Nakayama, H.Yamasaki and T.Saito : Sci. Rep., 8 (2018), Article number 5845.
- 5) J.Snoek, H.Larochelle and R.P. Adams : In Advances in Neural Information Processing Systems, (2012), 2951.
- 6) S.Iwazaki, Y.Inatsu and I.Takeuchi : AI & Statistics (AISTATS), (2021)
- 7) B.Bryan, R.C Nichol, C.R Genovese, J.Schneider, C.J Miller and L.Wasserman : Advances in Neural Information Processing Systems, 18 (2005), 163.
- 8) Y.Inatsu, M.Karasuyama, K.Inoue and I.Takeuchi : Neural Computation, 32 (2020) 12, 2486.

(2020年11月16日受付)