The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

パーシステント図に対するカーネルを用いたベイズ最適化

Bayesian Optimization with Kernels for Persistence Diagrams

白石 竜也 *1　山田 誠 *1*2　鹿島 久嗣 *1*2
Tatsuya Shiraishi　Makoto Yamada　Hisashi Kashima

*1京都大学 Kyoto University　*2理化学研究所 革新知能統合研究センター RIKEN Center for Advanced Intelligence Project

The graph structure optimization is a fundamental task in graph structured data analysis. In the graph structure optimization, objective function is usually an expensive-to-evaluate and black-box function. Therefore, we need to optimize the unknown function with as few evaluations of the function as possible. Bayesian optimization is one of methods that can handle this difficulty. However, in order to apply Bayesian optimization to graph structured data, we need to extract proper geometric information of the structure and measure similarity between the structures. In this paper, we utilize topological data analysis (TDA), which is recently applied in machine learning, to extract robust topological information from graph structured data. Experiments show that topological information extracted by TDA contributes to efficient search of the optimal structure compared with random search baseline.

1. はじめに

近年, グラフ構造データに対する機械学習手法の研究が盛んに行われている。グラフ構造最適化は最適な性質を持つグラフ構造を探索するタスクであり, グラフ構造データ分析のなかでも基本的で重要なタスクの一つである。例えば, あるタンパク質と最も強く結合する化合物の探索や, 交通量が最適となる道路ネットワークの探索などが挙げられる。

ここで最適化される目的関数, すなわち化合物の結合力や道路の交通量などといった性質は, 出力の評価に長時間の実験が必要な評価コストの高い関数であり, その形が与えられないブラックボックス関数であることが多い。そのため, 目的関数の評価回数を少なく抑えつつ, 未知の目的関数でも最適化できる手法が望ましい。この条件を満たす手法の一つにベイズ最適化がある。しかし, ベイズ最適化に関する研究はベクトルを入力とするものが多く, グラフ構造を扱うものは少ない。ベイズ最適化でグラフ構造を扱うには, その構造の幾何学的な特徴をうまく捉えた類似度を設計する必要がある。

一方で, 複雑な構造を持つデータのトポロジカルな特徴を抽出できる手法として, 位相的データ解析 (topological data analysis, TDA), 特にパーシステントホモロジーと呼ばれる手法が注目されている。この手法は距離空間上の点群から特徴抽出を行うもので, その結果は図 2 のようなパーシステント図 (persistence diagram, PD) と呼ばれる点群で表現される。

ベイズ最適化でグラフ構造を扱うための類似度設計において, TDA を利用する方法はこれまで行われてこなかった。そこで本研究では, PD に対するカーネルを用いることで, パーシステントホモロジーによって得られる情報を利用するベイズ最適化手法を提案する。また, パーシステントホモロジーでは一つのデータから複数種類の PD を抽出できるが, これら複数の PD の情報を同時に用いる手法は提案されてこなかった。本研究ではさらに, 複数のカーネルを線型結合で組み合わせることで, 複数の PD の情報を扱える手法を提案する。複数のデータを用いた実験を行い, ランダムに探索する場合と比べて, 提案手法によって効率的に探索が行えることを示す。

連絡先: 白石 竜也, 京都大学大学院情報学研究科知能情報学専攻，shiraishi.t@ml.ist.i.kyoto-u.ac.jp

2. 問題設定

入力空間 $\mathcal{X}$ から目的関数 $f : \mathcal{X} \to \mathbb{R}$ が最小となる点を探索する。本研究では $\mathcal{X}$ としてデータ集合 $\mathcal{D} = \{\boldsymbol{x}_i\}_{i \in I}$ を考え, この与えられたデータ集合から目的関数を最小化する点を探索する。ここで, $\boldsymbol{x}_i$ はそれ自体が点群やグラフ構造で表されるデータである。

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \in \mathcal{D}}{\operatorname{argmin}} \ f(\boldsymbol{x})$$

目的関数として評価コストの高い関数を想定しているので, できるだけ少ない評価回数で最適解を探索することが目標となる。ただし, 目的関数値は真の値 $f(\boldsymbol{x}_i)$ にガウス分布に従う独立な加法性ノイズ $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ が含まれた状態でのみ観測できるものとする。

3. ベイズ最適化

ベイズ最適化は評価コストの高い目的関数の最適化によく用いられる最適化手法の一つである。ベイズ最適化は繰り返し処理で構成されており, 各ステップはガウス過程による目的関数値の予測分布の計算と, 獲得関数を用いた次の探索点の選択の二つのステップからなる。

3.1 ガウス過程

ガウス過程は確率過程の一つであり, スカラーやベクトルに対するガウス分布を関数に対して一般化したものである。ベイズ最適化では目的関数 $f : \mathcal{X} \to \mathbb{R}$ をガウス過程でモデル化して予測分布の計算を行う。あるステップまでに入出力データ $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_t, y_t)\}$ が得られているとする。ただし, 出力 y_i として真の値 $f(\boldsymbol{x}_i)$ が観測されるとは限らず, ガウス分布に従う独立なノイズ $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ が含まれているとする。

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i$$

ガウス過程の定義より, このとき $f(\boldsymbol{x}_1), \cdots, f(\boldsymbol{x}_t)$ の同時確率分布はガウス分布となる。

$$(f(\boldsymbol{x}_1), \cdots, f(\boldsymbol{x}_t))^T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{K}) \tag{1}$$

ここで, $\boldsymbol{K}$ の各成分はカーネル関数を用いて $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ と表される。T は転置を表す記号である。このとき, データに含まれていない点 $\boldsymbol{x}_{t+1}$ に対する目的関数値 $f(\boldsymbol{x}_{t+1})$ の予測分布が計算できる。$f(\boldsymbol{x}_1), \cdots, f(\boldsymbol{x}_t), f(\boldsymbol{x}_{t+1})$ の同時確率分布も式 (1) と同様の形であることと, 出力にはノイズが含まれていることから, $f(\boldsymbol{x}_{t+1})$ の予測分布もまたガウス分布となり, その平均 $\mu(\boldsymbol{x}_{t+1})$ と共分散 $\sigma^2(\boldsymbol{x}_{t+1})$ は以下のようになる。

$$\mu(\boldsymbol{x}_{t+1}) = \boldsymbol{k}_{t+1}(\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{y}$$
$$\sigma^2(\boldsymbol{x}_{t+1}) = k(\boldsymbol{x}_{t+1}, \boldsymbol{x}_{t+1}) - \boldsymbol{k}_{t+1}(\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{k}_{t+1}^T$$

ここで, $\boldsymbol{k}_{t+1} = (k(\boldsymbol{x}_{t+1}, \boldsymbol{x}_1), \cdots, k(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t))$ 及び $\boldsymbol{y} = (y_1, \cdots, y_t)^T$ である。

3.2 獲得関数

ベイズ最適化では, ガウス過程を利用して計算された予測分布に基づいて, 獲得関数 $acq(\boldsymbol{x})$ が最大となる点を次に出力を評価する点として選択する。

$$\boldsymbol{x}_{t+1} = \underset{\boldsymbol{x} \in \mathcal{X}}{\operatorname{argmax}} \ acq(\boldsymbol{x})$$

獲得関数の選択においては, 既に得られている最良の値の周辺を調べる活用 (exploitation) と, 不確実性の高い部分を調べる探索 (exploration) とのバランスが重要となる。ここでは, よく使われている expected improvement (EI) について述べる。EI はある時点で得られている最良の値 y_{best} と, 予測される目的関数値 $f(\boldsymbol{x})$ との差の期待値を最大化する手法である。EI では獲得関数を以下で定義する。

$$\begin{aligned} & acq_{EI}(\boldsymbol{x}) \\ &= \mathbb{E}[\max\{0, y_{best} - f(\boldsymbol{x})\}] \\ &= \begin{cases} (y_{best} - \mu(\boldsymbol{x}))\Phi(Z) + \sigma(\boldsymbol{x})\phi(Z) & \sigma(\boldsymbol{x}) \neq 0 \\ 0 & \sigma(\boldsymbol{x}) = 0 \end{cases} \end{aligned}$$

ここで, $Z = \frac{y_{best} - \mu(\boldsymbol{x}_{t+1})}{\sigma(\boldsymbol{x}_{t+1})}$ であり, $\Phi(Z)$ と $\phi(Z)$ はそれぞれ標準正規分布の累積密度関数と確率密度関数である。

4. パーシステントホモロジー

位相的データ解析 (topological data analysis, TDA) ではトポロジーの観点から, 点群やグラフで表現されるデータの大まかな形に注目して分析を行う。トポロジーでは伸縮させて一致させられる図形同士は同じものとして扱われる。このような図形同士はホモトピー同値であるといわれ, 連結成分や穴の数が同じ図形となっている。ここでは, TDA 手法の一つであるパーシステントホモロジーについて直感的な説明を行う。

距離空間 (M, d) 上の点群 $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ をパーシステントホモロジーによって分析する。このとき, 各点を中心とする半径 r の領域をすべて足し合わせた領域 S_r を考える。

$$S_r = \bigcup_{i=1}^{N} \{\boldsymbol{x} \in M \mid d(\boldsymbol{x}, \boldsymbol{x}_i) \leq r\}$$

いくつかの r に対する S_r の例を図 1 に示す。r を大きくしていくと、領域同士が繋がって大小二つのリングが現れ, やがて小さいリングが潰れて無くなっていく様子が観察できる。パーシステントホモロジーでは, このとき現れる連結成分や穴などの幾何学的な構造について, それがいつ現れ, どのくらいの期間見られるのかに注目して分析を行う。

パーシステントホモロジーによって得られるトポロジカルな特徴は, パーシステント図 (persistence diagram, PD) と呼ばれる図として表現される。PD は r を大きくしていく過程で S_r に見られる幾何学的な構造について, その構造が現れたときの半径 b と潰れて無くなったときの半径 d のペア (b, d) をプロットしたものである。このとき注目する構造によって複数の PD が考えられ, 連結成分に注目した場合は 0 次の PD, リングに注目した場合は 1 次の PD などと呼ばれる。図 1 の点群に対する 0 次の PD と 1 次の PD を図 2 に示す。1 次の PD には大小のリングに対応する二つの点が見て取れる。小さい方のリングは birth と death の差が小さいことから分かるように, 対角線に近い点に対応している。同様に大きい方のリングは対角線から離れた点に対応している。このことから, 対角線の近くに分布している点はすぐに消えるノイズ的な構造を表しており, 対角線から離れて分布している点はより重要な構造を表していると考えられる。

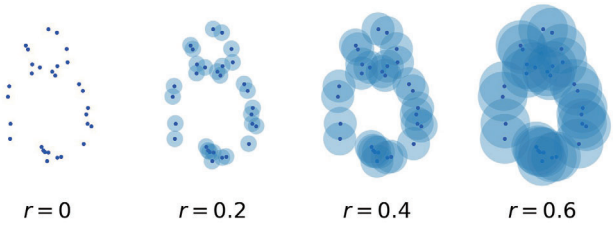

図 1: いくつかの r に対する S_r の例

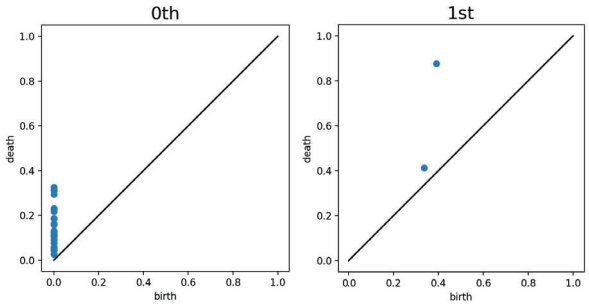

図 2: 図 1 の点群に対する 0 次の PD と 1 次の PD

5. 提案手法

本研究では, PD に対するカーネルを用いることで, パーシステントホモロジーによって得られる情報を利用したベイズ最適化を行う。また, 複数のカーネルを線型結合で組み合わせることによって, 一つのデータから得られる複数種類の PD の情報を同次に扱う方法を考える。

5.1 パーシステント図に対するカーネル

5.1.1 Persistence weighted Gaussian kernels

Persistence weighted Gaussian kernels (PWGK) [Kusano 16] は, PD を重み付き測度とみなしてカーネル平均埋め込みによって RKHS 上にベクトル化することで、通常の線形カーネル (PWGK-Linear) やガウスカーネル (PWGK-Gaussian) を考えられるようにする方法である。そ

れぞれは以下の式で表される。

$$k_L(D_i, D_j) = \sum_{\boldsymbol{x}\in D_i}\sum_{\boldsymbol{y}\in D_j} w(\boldsymbol{x})w(\boldsymbol{y})\exp\left(-\frac{||\boldsymbol{x}-\boldsymbol{y}||^2}{2\sigma^2}\right)$$

$$k_G(D_i, D_j) = \exp\left(-\frac{d(D_i,D_j)^2}{2\tau^2}\right)$$

ここで, $w(\boldsymbol{x})$ は各点を対角線からの距離に応じて重み付ける関数であり, [Kusano 16] では $w(\boldsymbol{x}) = \arctan(C\mathrm{pers}(\boldsymbol{x})^p)$ が使用されている。$\mathrm{pers}(\boldsymbol{x})$ は $\boldsymbol{x}$ の death と birth の差である。また, $d(D_i, D_j)^2 = k_L(D_i, D_i) + k_L(D_j, D_j) - 2k_L(D_i, D_j)$ と表される。

5.1.2 Persistence Fisher kernel

Persistence Fisher kernel (PFK) [Le 18] は, PD を平均だけが異なる正規分布の和とみなして、フィッシャー情報計量を用いて類似度を測る方法である。PFK ではパーシステント図 D_i, D_j を対角線に射影した $D_{i\Delta}, D_{j\Delta}$ を考え, D_i と D_j を比較する代わりに $D'_i = D_i \cup D_{j\Delta}$ と $D'_j = D_j \cup D_{i\Delta}$ を比較する。まず PD の各点を中心とする正規分布の和を考える。

$$\rho_{D'_i}(\boldsymbol{x}) = \frac{1}{Z}\sum_{\boldsymbol{\mu}\in D'_i}\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \sigma\boldsymbol{I})$$

ここで, Z は正規化定数である。そしてこの分布に対するフィッシャー情報計量を考える。

$$d_{FIM}(D_i, D_j) = \arccos\left(\int\sqrt{\rho_{D'_i}(\boldsymbol{x})\rho_{D'_j}(\boldsymbol{x})}d\boldsymbol{x}\right)$$

最終的なカーネルは以下の式で表される。

$$k_{PF}(D_i, D_j) = \exp\left(-td_{FIM}(D_i, D_j)\right)$$

5.2 複数カーネルの組み合わせ

パーシステントホモロジーで得られた複数の PD の情報を同時に扱うために, それぞれの PD から計算したカーネルを組み合わせて一つのカーネルを構成する。ここでは, 0 次の PD から計算したグラム行列 $\boldsymbol{K}_0$ と 1 次の PD から計算したグラム行列 $\boldsymbol{K}_1$ を線型結合によって組み合わせる。

$$\boldsymbol{K} = \alpha_0\boldsymbol{K}_0 + \alpha_1\boldsymbol{K}_1$$

係数 $\alpha_0, \alpha_1 > 0$ の学習方法として以下の方法を考える。

5.2.1 kernel target alignment

複数のカーネルを線形結合によって組み合わせる際に, 以下で定義される alignment を最大化するように係数を学習する方法が提案されている [Cortes 10]。

$$\rho(\boldsymbol{K}, \boldsymbol{Y}) = \frac{\langle\boldsymbol{K}_c, \boldsymbol{Y}_c\rangle_F}{||\boldsymbol{K}_c||_F||\boldsymbol{Y}_c||_F}$$

ここで, $\boldsymbol{Y} = \boldsymbol{y}\boldsymbol{y}^T$ 及び $\boldsymbol{K}_c(x, y) = \boldsymbol{K}(x, y) - \mathbb{E}_x[\boldsymbol{K}(x, y)] - \mathbb{E}_y[\boldsymbol{K}(x, y)] + \mathbb{E}[\boldsymbol{K}(x, y)]$ である。alignment の最大化は二次計画問題に帰着させることができる。ベイズ最適化では各ステップで $\boldsymbol{y}$ が更新されるので, 各ステップにおいて新たな出力が得られたタイミングで学習を行う。

5.2.2 最尤推定

ベイズ最適化では目的関数をガウス過程によってモデル化している。このとき, あるステップまでに出力データ $\boldsymbol{y} = (y_1, \cdots, y_t)^T$ が得られているとすると, $\boldsymbol{y}$ に対する対数尤度は以下のように計算できる。

$$\log p(\mathcal{D}|\boldsymbol{\alpha}) \propto -\frac{1}{2}\boldsymbol{y}^T(\boldsymbol{K} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{y} - \frac{1}{2}\log|\boldsymbol{K} + \sigma^2\boldsymbol{I}|$$

そこで係数を学習する方法として, 最尤推定によってこの対数尤度を最大化する方法を考える。対数尤度の最大化は勾配に基づく最適化手法などを用いて行うことができ, こちらも新たな出力が得られたタイミングで学習を行う。

6. 評価実験

6.1 データセット

人工データと 2 つの実データを用いて実験を行う。人工データの生成には [Hertzsch 07] で提案されている方法を用いる。$r > 0$ と初期点 $(x_0, y_0) \in [0, 1] \times [0, 1]$ から, 以下の式に従って点群 $\{(x_i, y_i)\}_{i=1}^N$ を生成する。

$$x_{n+1} = x_n + ry_n(1 - y_n) \mod 1$$

$$y_{n+1} = y_n + rx_{n+1}(1 - x_{n+1}) \mod 1$$

ここでは $N = 1000$ とした。$r = 2.0$ の場合と $r = 4.3$ の場合に生成される点群を図 3 に示す。この点群を一つのデータとし, 生成に用いた r をその目的関数値とする。データセットは r を $[2.0, 4.3]$ の範囲からランダムに 1000 個選んで生成する。

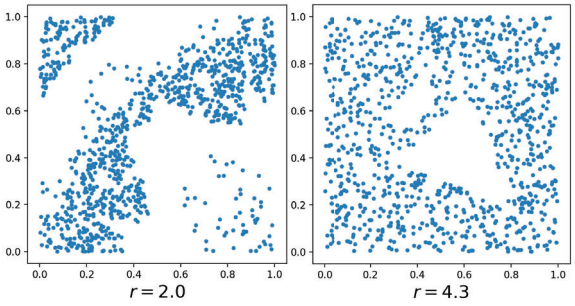

図 3: 人工データの例

実データには MoleculeNet [Wu 17] から ESOL データセットと FreeSolv データセットを使用する。どちらも比較的小さい化合物の性質に関するデータセットである。ESOL は水溶性に関するデータセットでサイズは 1128, FreeSolv は水和自由エネルギーに関するデータセットでサイズは 642 である。本研究では, 化合物を構成する原子の種類や結合は考慮せずに, 原子の 3 次元座標のみを用いて化合物を点群として扱う。

6.2 PD に対するカーネルによる効果

まずランダムに探索する場合と比べて, PD の情報によってどの程度効率的に探索できるようになるか検証する。ベイズ最適化では, ランダムに選んだ 10 個のデータを用いて最初の予測分布を計算する。カーネルとして PWGK-Linear, PWGK-Gaussian, PFK を用い, 獲得関数として EI を用いる。これらのカーネルは人工データについては 1 次の PD を用いて計算し, 実データについては 0 次の PD を用いて計算する。また, ハイパーパラメータの設定と近似計算を元論文 [Kusano 16][Le 18] に従ってそれぞれ行う。ベイズ最適化を 30 回実行したときの, 各ステップで得られている最小解の平均を図 4 に示す。ランダムに探索する場合と比べて, PD に対するカーネルを用いることで効率的に探索できていることが分かる。

6.3 複数のカーネルの組み合わせ

続いて複数種類の PD の情報を組み合わせることで, 一種類のみを用いる場合と比べて, どの程度効率的に探索できるようになるか検証する。0 次の PD から計算したカーネルを使用し

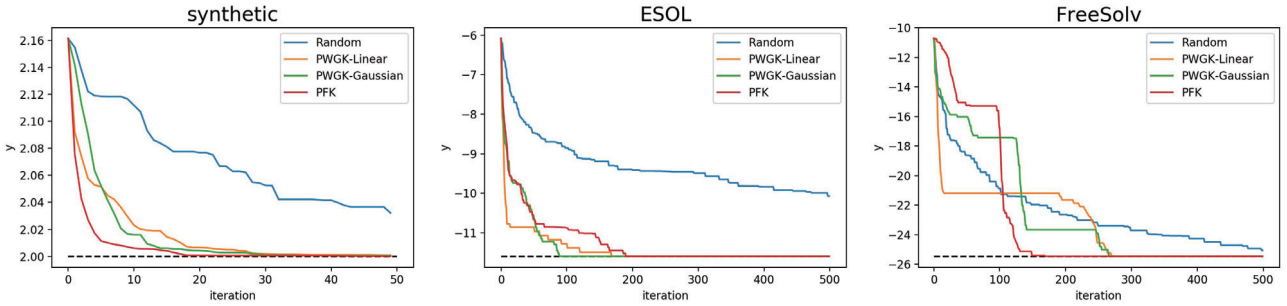

図 4: 各ステップにおいて得られている最小解の推移。黒い点線は探索対象である最適解を示している。

た場合 (0th), 1 次の PD から計算したカーネルを使用した場合 (1st), 両者を kernel target alignment によって組み合わせた場合 (align), 最尤推定によって組み合わせた場合 (MLE) とを比較する。ここでは評価指標として, 図 4 における収束曲線と黒い点線とで挟まれた部分の面積を用いる。各カーネルについて実験を行った結果を表 1 に示す。各値はランダムの場合が 1 になるようにスケーリングされている。多くの場合において, 最尤推定によって PD の情報を組み合わせることで, より効率的に探索できるようになることが分かる。

表 1: 複数カーネルの組み合わせによる実験結果

		Synthetic	ESOL	FreeSolv
Random		1.0000	1.0000	1.0000
PWGK -Linear	0th	0.1597	**0.0571**	0.6832
	1st	0.1551	0.3867	1.4169
	align	0.1664	0.3119	1.0350
	MLE	**0.0898**	0.1757	**0.5241**
PWGK -Gaussian	0th	0.1512	0.0763	0.8833
	1st	**0.1509**	0.4630	1.2399
	align	0.1618	0.2455	0.8862
	MLE	0.4308	**0.0560**	**0.5867**
PFK	0th	0.1172	0.1153	0.7685
	1st	**0.0730**	0.2544	0.6644
	align	0.0922	0.1195	0.8695
	MLE	0.2220	**0.0703**	**0.7640**

7. 関連研究

グラフ構造データに対するベイズ最適化に関する研究はいくつかあるが, 任意のグラフ構造に対するベイズ最適化のフレームワークとして, Graph Bayesian Optimization (GBO) が提案されている [Cui 18]。GBO では二つのカーネルの線形結合によってカーネルを構成している。一つはグラフの次数や各種中心性などを要素とするベクトルを用いた, ガウスカーネルなどのベクトルカーネルである。次数や中心性などの特徴量はグラフの幾何学的な特徴を表す値ではあるが, こうした明示的な特徴がそれをうまく反映できているとは限らない。そこで GBO では, ベクトルカーネルで表現されない情報を補うために, もう一つのカーネルとしてグラフカーネルを用いている。

8. むすび

本研究では, パーシステントホモロジーによって得られる複数のトポロジカルな情報を利用したベイズ最適化手法を提案した。PD に対するカーネルを用いた実験によって, PD の情報が効率的な探索に寄与していることと, 複数種類の PD の情報を組み合わせることでより効率的に探索できることを示した。

今後の課題として, 他の PD カーネルや獲得関数を用いた実験を試すことや, グラフカーネルを用いる GBO との比較を検討している。ベイズ最適化では, より実用的な問題設定として, 各イテレーションで一つずつデータを選択するのではなく, 複数の候補をまとめて選択することが考えられる。本研究では, 化合物のようなグラフ構造データを点群として扱ったが, グラフのノードやエッジに付加されている特徴を扱えるように拡張することが考えられる。また, 新たなグラフを生成する仕組みを追加することで, グラフを生成しながら未知のグラフ構造を探索できるようにすることも考えている。

参考文献

[Cortes 10] Cortes, C., Mohri, M., and Rostamizadeh, A.: Two-Stage Learning Kernel Algorithms, *In Proceedings of the 27th International Conference on Machine Learning* (2010)

[Cui 18] Cui, J. and Yang, B.: Graph Bayesian Optimization: Algorithms, Evaluations and Applications, *arXiv preprint arXiv:1805.01157* (2018)

[Hertzsch 07] Hertzsch, J.-M., Sturman, R., and Wiggins, S.: DNA microarrays: design principles for maximizing ergodic, chaotic mixing, *Small*, Vol. 3, pp. 202–218 (2007)

[Kusano 16] Kusano, G., Fukumizu, K., and Hiraoka, Y.: Persistence weighted Gaussian kernel for topological data analysis, *In Proceedings of the 33rd International Conference on Machine Learning*, pp. 2004–2013 (2016)

[Le 18] Le, T. and Yamada, M.: Persistence Fisher Kernel: A Riemannian Manifold Kernel for Persistence Diagrams, *Advances in Neural Information Processing Systems*, pp. 10027–10038 (2018)

[Wu 17] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V.: MoleculeNet: A Benchmark for Molecular Machine Learning, *arXiv preprint arXiv:1703.00564* (2017)