講演番号：97

# h-outlier r-gathering 問題

h-outlier r-gather problem

坪井　誠　　中野　眞一

Makoto TSUBOI　　Shinichi NAKANO

群馬大学大学院理工学府

Graduate School of Science and Technology, Gunma University

## 1. はじめに

大量のデータ(時系列データなど)の特徴をおおまかに捉えて俯瞰することが多くの分野で必要とされている。n 点の集合 P を k 段のステップ関数(図1参照.)で近似するアルゴリズムがある [1]. [1]では, 各ステップに対応する点の個数に制限はないため, ステップ関数が点集合 P をよく俯瞰しているとは必ずしもいえない.(図2参照.)また, 外れ値を大量のデータから削除することが望ましいことがある.

本文では, あたえられた n 点の集合 P から h 個の点を除去した点の集合を, 各ステップが r 個以上の点に対応し, かつ, max コストが最小となるステップ関数で近似する問題をあつかう.

## 2. 定義

ステップ関数とは, 定義域がいくつかの区間からなり, 定義域の各区間の値が定数となる関数である. 直感的には, 階段状の関数である. 例を図 1 に示す. あるステップに対応する各点の y 座標とステップの y 座標の差の最大値を, そのステップの max コストとする. また, ステップ関数の各ステップの max コストの最大値をそのステップ関数の max コストとする. n 点の集合 P から, h 個の点を除去した点の集合を, 各ステップが r 個以上の点に対応し, かつ, max コストが最小となるステップ関数で近似する問題を, h-outlier r-gathering 問題とよぶ.

## 3. アルゴリズム

本文ではこの問題を解く動的計画法によるアルゴリズムを設計する. 左から i 個の点の集合 $P_i \subseteq P$ から, ちょうど h'(≦h)個の点を除去した点の集合を, 各ステップが r 個以上の点に対応し, かつ, max コストが最小となるステップ関数で近似する問題を, 部分問題 P(i, h') とする. この問題を解くアルゴリズムを設計しよう. 3つの場合に分けて考えよう.

**場合1** $i < r+h'$ のとき. 解なし

**場合2** $r+h' < i < 2r+h'-1$ のとき. ステップはちょうど1段である.

ステップに対応する各点の y 座標の最大値を $y_{max}$, 最小値を $y_{min}$ とすると, ステップの y 座標が, $(y_{max}-y_{min})/2$ のとき, ステップの max コストは最小となる. O((r+h')log (r+h')+h')時間で部分問題 P(i, h') を解くことができる.

**場合3** $i > 2r+h'-1$ のとき. ステップは2段以上である.

一番右のステップと, それ以外のステップの集合の組合せで解は構成される. 一番右のステップの外れ値 h''= 0,1,…,h'のいずれかの個数含み, かつ, 対応する点を r+h'', r+h''-1,…, 2r+h''-1 個のいずれかの個数含む.

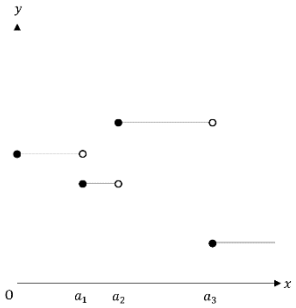


図 1　4段のステップ関数の例

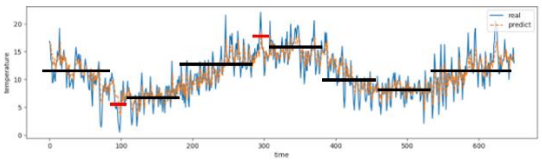


図 2　ステップ関数による近似の例

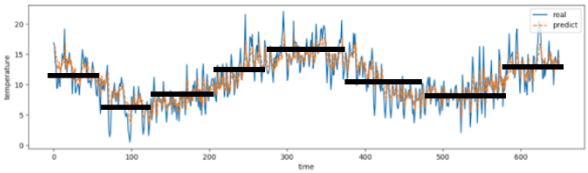


図 3　各ステップが r 個以上の点に対応するステップ関数による近似の例

これらの各場合について, 適切な部分問題を組み合わせて得られる問題 P(i, h') の解の候補のうち, 最も max コストの小さいものを計算する.

単純な動的計画法では, この問題を解くのに $O(nh^2r((r+h)\log(r+h)+h))$時間かかる. 本研究ではこれを $O((r+h)^2h+(r+h)h^2n)$時間に高速化する. 主なアイディアは平衡木 [2]を利用することである.(詳細略)

## 参考文献

[1] Hervé Fournier, et al., Fitting a Step Function to a Point Set., Algorithmica, 60:95-109 (2011).

[2] Thomas H. Cormen, et al., Introduction to Algorithms, MIT Press (2009).