様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月　 ２ 日現在

機関番号：１２６０１
研究種目：基盤研究(C)（一般）
研究期間：2018～2022
課題番号：１８Ｋ１１４２６

研究課題名（和文）高次元特徴空間の大規模データから支配的境界集合を抽出する研究

研究課題名（英文）Extraction of dominant boundary set from high dimensional data

研究代表者
稲葉　真理（inaba, Mary）

東京大学・大学院情報理工学系研究科・准教授

研究者番号：６０２８２７１１

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：大規模データを用いた学習・予測・探索の性能を向上させるため、オリジナルデータからアプリケーショ
ンの目的に合致したサンプルデータを抽出するための手法の提案を行った。具体的には、ランダムサンプリングでは困難な裾野データや、ある特徴量で最大値をとるといった「際立った特徴を持つデータ」を、サンプル集合の中に含めることを目標とし幾何的性質を利用しながらも、特徴空間においてデータ集合の凸包を求めると高次元での計算コストが非常に高いため、本研究では、スカイライン問題、および、高速にスカイライン問題を解くためのBJR-tree 構造を用いたデータサンプリングを提案し、低中次元における実装実験を行った。

研究成果の学術的意義や社会的意義
BJR-tree 構造を用いて、pseudo-skyline 問題を効率よく解くことで、サンプル抽出することにより、「際立った特徴を持つデータ」をこぼすことなく、サンプル集合を得ることができる。この手法を、TCP/IP輻輳制御問題および、アプリケーション実行時に利用されるアドレス検出の強化学習実験で検証した結果、低中次元の特徴空間においては、概ね、予想通りの結果を得ることができたたが、高次元化すると、psuedo-skyline 集合が爆発的に増大してしまうため、多様性を用いた大規模データの探索問題に問題範囲を広げ、幾何構造に重ねる形でグラフ構造（本研究ではラティス）を用いた提案も行った。

研究成果の概要（英文）：In order to improve the performance of learning, prediction, and search using large-scale data, we proposed a method for extracting good sample data set. For example, random sampling rarely samples the "data with outstanding features" such as the maximum value of a certain feature value, and naive geometric approach to get such kind of data is to compute convex hull in the feature space, whose computing cost is extremely high especially in the high dimensional space. To tackle with this problem, we utilize BJR-tree structure, which is originally invented to solve the skyline problem using dominance relationship between a pair of data in the feature space. Roughly speaking, this approach is, to convert the geometric problem into graph (or, tree) problem, and the computational cost is not high comparing with computing convex hull in the high dimensional space. Experimental result shows that this approach is good for the low and the middle dimensional space.

研究分野：アルゴリズム

キーワード：最適化問題　幾何構造を利用する最適化　グラフ構造を利用する最適化　スカイライン問題　SAT ソルバ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

大規模データを用いた学習・予測・探索では、オリジナルデータからサンプルデータを抽出し操作を行うことが多い。ここでは、抽出されたサンプル集合の性質が、学習・予測・探索の効率に対し大きな影響を与えることが知られており、アプリケーションの目的に合致したサンプル集合を効率よく求める手法が求められている。ランダムサンプリングは、非常に高速に行えるため、しばしば、用いられているが、裾野データが抽出される確率は低く、たとえば、ある特徴で最大値をとるといった「際立った特徴を持つデータ」は普通はサンプルされないため、オリジナルのデータ集合が持つ多様性が失われてしまうことが多い。一方、特徴空間を幾何的に捉え、たとえば、多次元空間にデータをマップし、その凸包を求めれば「特徴が際立ったデータ」を抽出することができる。しかしながら、現実のデータには、たとえば、観測データにおけるノイズのような、本来は、含まれてはいけないないような外れ値が含まれていることがあり、外れ値を中心とするデータ集合を集めると、それはそれでミスリーディングとなってしまう。また、実際に計算を行うと、高次元で凸包の計算コストが非常に高いことも知られており、特徴空間が高次元になる場合に、幾何的な手法を使うと、計算効率が悪くなることが懸念される。

２．研究の目的

本研究では、大規模データを用いた学習・予測・探索を行う際に、上述したような「際立った特徴を持つデータ」がこぼれないようなサンプル抽出を行うことで、学習・予測・探索効率を上げることを目的とする一方で、観測ノイズのような、外れ値データに引きずられないサンプル集合を効率よく求めることを目的とする。直感的には、特徴空間で、幾何的に捉えた時、他のデータの内分点として表されるデータの抽出確率を、ある程度下げ、データ集合のうち、集合の外側に位置するようなデータを抽出することで、際立った特徴を持つデータの取りこぼすことがない効率的なサンプリング手法を確立することを目的とする一方で、外れ値データには、さほど引きずられないことを目標とする。具体的には、幾何的性質を持つデータに対し、２点間の位置的な支配関係を考え、他点からの非被支配データの集合は、スカイラインと呼ばれており、スカイラインを効率的に求めるための色々なアルゴリズムが提案されてきている。この支配・被支配関係に着目し、被支配される支配データが少ないデータが「際立った特徴を持つデータ」と考えることで、スカイライン抽出アルゴリズムを拡張することで、際立ったデータを取りこぼさない効率的なサンプルデータ抽出手法を提案・検証することを目的とする。

３．研究の方法

幾何的性質を持つデータに対し、２点間の位置的な支配関係を考え、他点からの非被支配データの集合は、スカイラインと呼ばれており、スカイラインを効率的に求める

ための色々なアルゴリズムが提案されてきている。このような手法の中に、全点間の支配関係を用いたグラフを構成すると、半順序関係を表すグラフが構成でき、この最上位にあたるデータは、被支配関係がある点が存在しない点となるため、最上位のデータ集合が、スカイラインとなることが知られている。このグラフに対して、すべての点を支配する点を　ルートとして加えた上で、このグラフのサブグラフに当たるBJR-tree を用いたスカイライン抽出アルゴリズムが提案されている。このBJR−tree構造を用いて、上から１段目に当たるデータ集合が、スカイライン、上からk 段以内にある被支配される支配点が少ない集合を　pseudo-skyline と新たに定義し、BJR-tree アルゴリズムを拡張し、pseudo-skyline を求めることで、効率よく、データ集合の外郭に位置するデータ集合を求める。被支配される支配店集合がが少ないデータを抽出することで、際立った特徴をもつデータを取りこぼさない、サンプル集合を求めることができると仮定し、pseudo-skyline を求める効率的なアルゴリズムの提案・実装実験を行う。また、FPGA を用いたハードウェア化によって高速化も行う。しかしながら、本研究で実験を重ねるうちに、pseudo-skyline 問題は、次元を上げることにより、pseudo skyline 集合が急激に増大してしまうことが多いことが確認されたたため、」次元削減を試みる一方で、似たような半順序関係が成立するグラフを用いた他の問題への応用も考えることとした。その一つとして、SAT 問題に多項式時間還元できるアプリケーション最適化問題で、ラティス構造をグラフ的に捉えることで、、問題のdiversity を SSI という指標で計量し、効率よく最適化を行う応用を、当初の計画に対し、追加し行った。

４．研究成果

これまでの分割統治法や空間的情報を使う手法は高次元データに適していなかった。我々の提案する、エントリ間の関係を有向辺で表現する新たな木構造BJR 木では、深い頂点の探索を遅延させることで計算量を削減する。また、BJR 木を用いたハードウェアアルゴリズムLow-latency Skyline Computation Accelerator（LSCA）は、エントリ関係の判定処理を並列化し、待機時間に深く木探索を行うことで高速化する。BJR 木は既存のLookOut アルゴリズムより約70 倍高速であった。またLSCAをFPGA に実装した結果、ソフトウェア実装と比較して、乱数生成データに対し2.5 倍から4.4 倍、実データに対し1.7～35 倍高速であった。以上を含むアクセラレータの設計を通して我々が確立した設計の方法論では、複数の粒度で計算を並列化するために並列性の分類とパターン化を行い、アプリケーション指向のデータレイアウトを行う。我々は、具体的な計算処理において、動作合成で生成された回路と我々の方法論に基づいて設計された回路を比較し、我々の方法論の効果を示した。また、検証のために、この手法をTCP/IP輻輳制御問題および、アプリケーション実行時に利用されるアドレス検出の強化学習実験を行った。低中次元の特徴空間においては、概ね予想通りの結果を得ることができたたが、高次元化すると、skyline 集合、および、psuedo-skyline 集合が爆発的に増大してしまうことがわかった。これは、直感的には、次元が上がることで、支配・被支配関係が薄まることを意味しており、対応としては、(1)次元削減、及び、(2)支配・被支配関係の緩和によるグラフの再定義が考えられるが、現在、どちらの手法についても、さほど良い結果は得られていない。そのため、似たような半順序関係が成立するグラフを用いた他の問題への応用も考えることとした。その一つとして、SAT

問題に多項式時間還元できるアプリケーション最適化問題で、ラティス構造をグラフ的に捉えることで、問題の diversity を SSI という指標で計量し、効率よく最適化を行う応用を、当初の計画に対し、追加し行った。

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Koizumi Kenichi、Eades Peter、Hiraki Kei、Inaba Mary	４．巻 7
２．論文標題 BJR-tree: fast skyline computation algorithm using dominance relation-based tree structure	５．発行年 2018年
３．雑誌名 International Journal of Data Science and Analytics	６．最初と最後の頁 17～34
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s41060-018-0098-x	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Nitta Nao、Sugimura Takeaki、Isozaki Akihiro、Mikami Hideharu、Hiraki Kei、Sakuma Shinya、Iino Takanori、Arai Fumihito、Endo Taichiro、Fujiwaki Yasuhiro、Fukuzawa Hideya、Hase Misa、Hayakawa Takeshi、Hiramatsu Kotaro、Hoshino Yu、Inaba Mary、...、Ozeki Yasuyuki、Goda Keisuke	４．巻 175
２．論文標題 Intelligent Image-Activated Cell Sorting	５．発行年 2018年
３．雑誌名 Cell	６．最初と最後の頁 266～276.e13
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.cell.2018.08.028	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 下見淳一
２．発表標題 シングルサーバを用いた100 Gbpsネットワークでのセキュアファイル転送
３．学会等名 IA 研究会
４．発表年 2019年～2020年

１．発表者名 濱崎福平
２．発表標題 リソース制約条件を考慮した多目的最適化による高位合成
３．学会等名 CPSY 研究会
４．発表年 2019年～2020年

１．発表者名 Koizumi Kenichi、Hiraki Kei、Inaba Mary
２．発表標題 Application of a fast skyline computation algorithm for serendipitous searching problems
３．学会等名 SPIE BIOS（国際学会）
４．発表年 2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関