様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　５月２５日現在

機関番号：10101
研究種目：基盤研究(C)
研究期間：2009～2011
課題番号：21500128
研究課題名（和文）　個々のデータに依存した問題の複雑さに関する研究

研究課題名（英文）　A Study on Data-Dependent Complexities

研究代表者
中村　篤祥（NAKAMURA ATSUYOSHI）
北海道大学・大学院情報科学研究科・准教授
研究者番号：50344487

研究成果の概要（和文）：
データ依存の問題の複雑さ指標として，Sperner 族のデータ依存の VC 次元を提案し，超矩形サブクラス問題において，データ依存の VC 次元が低いデータに対し，高速に動作するアルゴリズムを開発し，実データを用いて有効性を検証した．また，文字列の連続繰返し構造に対する複雑さの指標として，繰返し表現文字列の最小サイズを提案し，実際に最小繰返し表現文字列を高速に求めるアルゴリズムを開発し，DNA の繰返し構造を分析した．

研究成果の概要（英文）：
As a data-dependent complexity measure, we proposed data-dependent VC dimension of Sperner family, and in hyper-rectangle subclass problem, we developed a fast algorithm for datasets with small such VC dimension. We also empirically showed efficiency of the algorithm using real data. Furthermore, as a complexity measure of contiguous repetition structure of a string, we proposed the size of a minimum repetition representation string (MRRS) for a given string. We developed a fast algorithm for constructing an MRRS and analyzed the repetition structure of DNA sequences using the algorithm.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,300,000	390,000	1,690,000
2010 年度	800,000	240,000	1,040,000
2011 年度	1,100,000	330,000	1,430,000
年度			
年度			
総　計	3,200,000	960,000	4,160,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：知識発見とデータマイニング，データ依存，問題の複雑さ，機械学習，シーケンス，超矩形，アルゴリズム

１．研究開始当初の背景

世の中のほとんどの問題は時間計算量において困難性が証明されているにもかかわらず，実際の問題例は意外と少ない計算量で解ける場合が多い．学習可能性にしても理論的に導かれる必要事例数が非常に多い場合でも，実際の問題例では少ない事例から高精度の予測ができる場合が少なくない．

この現象を説明する１つの仮設は，「データ依存の問題の複雑さというものがあって，現実のデータに対してはその複雑さが小さい」というものである．

データ依存の複雑さを表す指標で既に存在するものとして，サポートベクターマシン(SVM)におけるマージンがあげられる．イギリスの J. Showe-Taylor らによって，マージンが大きいほど予測性能が良い可能性が高いことが理論的に示されている．

２．研究の目的

本研究では，解くのが難しいと言われているいくつかの問題に対して，実際の問題例では何故，意外と簡単に解けてしまうのかを，データ依存の複雑さを表す何らかの指標を使って説明することを試みる．また，開発した指標で単純な問題を，より効率的に解くアルゴリズムを考案することを目指す．

特に以下の問題を中心に解析を行う．

(1)Sperner 族からなる概念クラスに属する概念の列挙問題

極大頻出アイテムセットの列挙は，計算困難な問題であるにもかかわらず，高速なアルゴリズムが開発されている．Sperner 族からなる概念クラスに属する概念の列挙は、この問題を一般化した問題.

(2)シーケンスデータに関する問題

時系列データや文字列データなどに関連する問題でそれぞれのシーケンスの複雑さが求解の難しさに影響していると思われる問題.

３．研究の方法

(1)Sperner 族からなる概念クラスに属する概念の列挙問題

我々は，2008 年の論文にて，この問題に対するデータ依存の複雑さを，概念クラスの intersection closure の VC 次元で測る方法を提案した.

極大頻出アイテムセットの列挙問題の実データでデータ依存の VC 次元を測ったところ，よく使われている列挙アルゴリズムによる計算時間とこの複雑さの指標がある程度一致していることが分かった．更に，この一般化された問題の他のインスタンスである超矩形サブクラス問題に関しても，実データを使ってデータ依存の VC 次元を測ったところ，意外に小さいものがあることがわかった.

超矩形サブクラス問題は，高速な列挙アルゴリズムが開発されていないが，極大頻出アイテムセットの列挙問題は，最近のデータマイニングブームで高速なアルゴリズムが開発されている.

そこで，超矩形サブクラス問題用の高速アルゴリズムを開発し，今まで解けなかった中規模データに対してまで，この問題が解けるようにする.

(2)シーケンスデータに関する問題

文字列の繰り返し構造の複雑さを表す指標を考案し，実際の DNA データなどを用いて有効性を検証する.

オンライン学習問題は，シーケンスでデータが与えられる学習問題である．オンライン学習問題に関し，データ依存の学習の難しさに関する指標を考え，その指標を用いて分析を行う.

４．研究成果

(1)Sperner 族からなる概念クラスに属する概念の列挙問題

①アルゴリズム LCMmax.R および MRF の開発 ［雑誌論文①］

超矩形サブクラス問題を解くアルゴリズム LCMmax.R を開発した．このアルゴリズムは，多次元実数空間上の与えられた正負ラベル付きサンプルに対し、負例を含まない軸に平行な超矩形で囲まれる正例の集合を列挙するものである.

このアルゴリズムは，極大頻出アイテムセットを列挙する問題と超矩形サブクラス問題の類似性に着眼し，極大頻出アイテムセット列挙の高速アルゴリズム LCMmax を超矩形サブクラス問題用に変換したものである．我々は超矩形サブクラス問題の特性を考慮した改良も行い，LCMmax.R というアルゴリズムを開発した.

LCMmax.R では，単純に変換したアルゴリズムより、データによっては 10 倍程度高速化が実現されている。また、このアルゴリズムを最大のものしか求めないように改造したアルゴリズム MRF を繰り返し使うことにより行う，超矩形による貪欲被覆アルゴリズムは，中規模の実データに適用可能なほど高速であることを実験により確かめた。超矩形による貪欲被覆はクラス分類やデータマイニングで有用であり、計算量の問題があって現在使われていないが、本研究の成果による高速化により実用化に一歩近づいたといえる。

(2)シーケンスデータに関する問題

①最小繰返し表現文字列

計算アルゴリズムの開発[雑誌論文③]

１つの文字列内には複数の連続的な繰り返しが存在する。DNA シーケンスなどに存在するそのような連続的な繰り返しはタンデムリピートと呼ばれ、遺伝的疾患に関係があることが知られている。１つの文字列の中に繰り返しは多く存在し、それらは互いに重なっていたり入れ子になっていたりする。入れ子になっているか、互いに重なっていない複数のタンデムリピートは、１つの「繰返し表現文字列」として表現できる。どのタンデムリピートを表現するかにより、同じ文字列が複数の（サイズの異なる）繰

返し表現文字列で表現できる。その最小サイズを、その文字列の繰返し構造の複雑さを表す指標として提案した。この指標において複雑度の低い文字列は、連続する繰返し構造という観点において、規則正しいことを意味する。

我々は、与えられた文字列に対する最小繰返し表現文字列を求める効率的なアルゴリズムを考案し、実際の複数の生物種の DNA シーケンスの繰返し構造の複雑度を計算した。その結果、元々の文字列長で正規化した複雑度には生物種に固有の値があり、またタンデムリピートの密度が同じでも正規化複雑度が異なる生物種が存在することがわかった。（図１参照。）

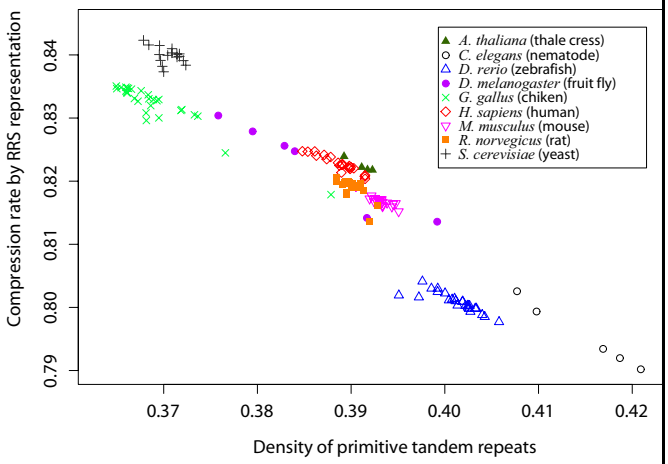

図１：*９種の主な生物種の繰返し構造の正規化複雑度。小さいほど単純な繰返しからなっていることを表す。横軸はタンデムリピート密度。*

②散在繰返し近似パターン列挙アルゴリズムの開発[雑誌論文②,学会発表①]

①の研究とは異なり，連続する繰り返し構造ではなく，散在する繰り返し構造に着目した。DNA シーケンスなどに散在する繰り返しはレトロトランスポゾンの転移によりできたものと考えられており、生物の進化と密接な関係がある。散在する繰り返しの近似パターンを列挙する問題のデータ依存の複雑さに関して考察を行った。新たな複雑さ指標による特徴付けはできなかったが、局所最適な出現が頻出する近似パターンの列挙を、ギャップ制約の下で $O(n^2)$のメモリで行うアルゴリズムを開発した。ただし、n はシーケンス全体の長さとする。さらに、実際のヒトゲノムから長さが 100 以上で 5000 万塩基あたり 100 回以上（ギャップ１の制約下で）局所最適出現するパターンの列挙に成功した。

③アルゴリズム Exp3.M の開発

[雑誌論文④,学会発表②,③]

オンライン学習問題の１つである，多腕バンディット問題において，１度に k 腕選択する問題に対するアルゴリズム Exp3.M を開発し，リグレット評価を行った. Exp3.M は Auer らが開発した Exp3 の拡張であるが，リグレットに関しても Auer らの結果を拡張する上界を得た. リグレットは，新たに提案した指標ではないが，ある条件を満たすベストのアルゴリズムの性能との比較であり，データによってはベストアルゴリズムでも性能が悪い場合もあることを考えると，ある程度データ依存の問題の難易度を考慮した評価といえる.

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 8 件）

①K. Ouchi, A. Nakamura, M. Kudo, Efficient construction and usefulness of hyper-rectangle greedy covers, Proceedings of GrC 2011, pp. 533-538, 2011, 査読有.

②Nakamura, M. Kudo, Packing Alignment: Alignment for Sequences of Various Length Events, LNCS 6635 (Proceedings of PAKDD 2011), pp. 234-245, 2011, 査読有.

③Nakamura, T. Saito, I. Takigawa, H. Mamitsuka, M. Kudo, Algorithms for Finding a Minimum Repetition Representation of a String, LNCS 6393 (Proceedings of SPIRE 2010), pp. 185-190, 2010, 査読有.

④T. Uchiya, A. Nakamura, M. Kudo, lgorithms for Adversarial Bandit Problems with Multiple Plays, LNCS 6331 (Proceedings of ALT2010), pp. 375-389, 2010, 査読有.

〔学会発表〕（計 5 件）

①中村, DNA シーケンスからの近似頻出パターンの発見, 人工知能学会 第 85 回 人工知能基本問題研究会, 2012 年 2 月 2 日, 下呂交流会館（岐阜県）.

②中村, バンディットの理論と応用, 第 14 回情報論的学習理論ワークショップ (IBIS2011)(招待講演), 2011 年 11 月 11 日, 奈良女子大学 講堂（奈良県）.

③中村, Capped Hedge Algorithm に関する一考察, 人工知能学会 第 82 回 人工知能基本問題研究会, 2011 年 8 月 4 日, 釧路工業高等専門学校（北海道）.

〔図書〕（計０件）

〔産業財産権〕
○出願状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
特になし.

６．研究組織
(1)研究代表者
中村　篤祥（NAKAMURA ATSUYOSHI）
北海道大学・大学院情報科学研究科・准教授
研究者番号：50344487

(2)研究分担者
工藤　峰一（KUDO MINEICHI）
北海道大学・大学院情報科学研究科・教授
研究者番号：60205101
外山　淳（TOYAMA JUN）
北海道大学・大学院情報科学研究科・助教
研究者番号：60197960

(3)連携研究者
（　　）

研究者番号：