様式 C-19

科学研究費補助金研究成果報告書

平成２３年　５月３１日現在

機関番号：12601
研究種目：若手研究（B）
研究期間：2008～2010
課題番号：20700135
研究課題名（和文）　内包に基づくカーネルによる構造データ学習と知識発見

研究課題名（英文）　Learning and discovering knowledge for Structured data by kernel function based on intention

研究代表者
土井　晃一郎（DOI KOICHIRO）
東京大学・大学院新領域創成科学研究科・特任講師
研究者番号：10345126

研究成果の概要（和文）：カーネル関数を使用した学習手法は盛んに研究されてきた。本研究では我々が提案したカーネル関数の新たな設計手法である内包カーネル関数を RNA 配列などの具体的なデータに適用した。また、木構造データに対する圧縮を利用した知識発見手法を根付き順序木、根付き無順序木の場合に提案し、半構造化文書に対して実験を行い、効率的に知識発見が行えることを示した。

研究成果の概要（英文）：Learning by using kernel functions has been actively researched. This study applies intentional kernel function to RNA sequence or other real data. We present a new method for finding frequent patterns from tree structured data in the case of rooted ordered tree and rooted unordered tree. We made experiments of our proposed method with some semi-structured data and show efficiency of our method.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	1,100,000	330,000	1,430,000
2009 年度	800,000	240,000	1,040,000
2010 年度	800,000	240,000	1,040,000
年度			
年度			
総　計	2,700,000	810,000	3,510,000

研究分野：アルゴリズム、バイオインフォマティクス
科研費の分科・細目：情報学・知能情報学
キーワード：機械学習、カーネル関数

１．研究開始当初の背景

サポートベクトルマシンをはじめとするカーネル関数を使用した学習手法(以下、カーネル法)は国内外において盛んに研究されている。カーネル法とは訓練データである正例、負例を空間上の点として入力され、別の高次元空間上の内積をカーネル関数で与えることにより高次元空間上における線形識別関数を効率的に求める手法である。空間上の点を扱うことから、このカーネル法は数値属性を持つデータを対象としていた。

一方、生命情報学における遺伝子配列データや自然言語文書データ，Web ページに利用される半構造化データなどの離散的で構造を持つデータ(以下，構造データとよぶ)が爆発的に増えてきていることから，その大規模なデータの活用が求められている。このことを背景にして、構造データを対象としたカーネル関数が考えられてきている。つまり、構造データの持つ様々な属性を数値で表し、ある特徴空間上の点として表現するカーネル関数を設計する研究がさかんに行われてきている。

構造データに対するカーネル関数の多く

は畳み込みカーネルの考え方に基づいている。畳み込みカーネルでは，構造データを部分構造の集まりと考え，データの部分構造同士のカーネル関数によってより大きな構造データのカーネル関数を再帰的に定義している．このことは,1つのデータを，その中に含まれるすべての部分構造を特徴量とするような無限次元の特徴空間上の1点に射影することでカーネル関数を定義していることになる．

しかし、構造データの捕らえ方はこれだけではない。成分が0と1だけからなる有限次元のブール値ベクトルの形をしたデータに対する DNF カーネルを Khardon(米国タフツ大学)と佐土原(産業総合研究所)が独立に発見している。DNF カーネルは，任意のブール関数が選言標準形で記述できることを利用し，与えられた命題変数から構成可能なすべての連言を特徴量とする高次元空間への射影を行う。しかし、ブール値ベクトルの部分構造として「連言」という論理式は入っていないことである．したがって，DNF カーネルは畳み込みカーネルとは全く異質のカーネルである。

データの内包とは別の言葉で言うとその構造データに対する説明のことであり、部分構造ではなくデータの概形や枠組みがデータの特徴であるという考え方である。例で説明すると、人の顔を目や鼻といった構成部品からではなく、目や鼻の位置関係といった概形、輪郭で似ている、似ていないを判断しようという考え方である。

２．研究の目的

本研究の目的は、具体的な遺伝子データ、XML データなどに適用できるカーネル関数を内包に基づくカーネルの考え方に基づいて設計、実験を行うことにより、この内包に基づくカーネルの考え方、手法の実データに対する有用性を明らかにすることである。今までの研究で、一階述語論理の項に対しては、この考え方が適用できること、項の間で成り立つ性質を利用してどのように計算を行うが効率的なのかを明らかにしてきているが、実データに対してこの考え方を適用するためには、何を具体的に特徴量とするのか、そのカーネルをどのように計算するかなど明らかにしなくてはいけない課題がまだ多数残っている。それらの課題をクリアして、どのようなデータのどのような概念がこの内包に基づくカーネル関数でうまく学習できるか、あるいは、既存の畳み込みカーネルの方が有用なのかといった諸性質を明らかにし、このことによって、この内包という考え方に基づくカーネル関数が実際のデータに対してどの程度有用なのかを明らかにしていくことを目的としている。

３．研究の方法

(1) 実データに対するカーネル関数の設計

遺伝子配列データや自然言語文書データ、Web ページに利用される半構造化データなど様々な構造データが存在するために、はじめに着手する適用例について検討を行う。申請者がこれまで生命情報学の研究を主に行ってきて、ある程度の基礎知識のあることから

・RNA の機能分類　　microRNA などのたんぱく質に翻訳されな RNA の働きの重要性が明らかになってきたことにより、RNA の未知の機能を明らかにする研究が注目されてきている。

・遺伝子ネットワーク　シグナル伝達や代謝パスウェイ、遺伝子間相互作用データ、遺伝子ネットワークにおける分類

などに対して適用することを考えている。この研究では非コード RNA を機能で分類する問題に対して、二次構造を文脈自由文法で表現することによって内包カーネルの考え方を適用している。この研究を更に発展させることも計画している。　また、半構造化文書は意味的な階層構造がそのままタグの木構造として表現されているので内包カーネルの考え方を適用することが比較的容易であるかもしれないと考えている。この段階においては、関連する研究者に話を聞いて、慎重に検討していくことが重要であると考えている。

このカーネル関数の設計には、構造データの概念がしっかりと表現できているかとともに、計算する上での効率も求められなくてはならない。実際、我々は先行研究において、特徴量をうまく分割して計算する手法を提案している。実データに対して適用していく上において、こういった実データの特徴を生かした計算の工夫は必要になってくると思われる。また、実データに対して適用していくには、説明とは何かということに対してすべての特徴量を数え上げるのではなく、必要な特徴量だけを使用してカーネル関数を構築するといった視点も必要になるかもしれない。こういった個々のデータの特徴をうまく捉えてカーネル関数の設計を行っていく。先行研究における RNA クラス分類に対するカーネル関数の研究を発展させるとともに，他の XML データなどに対するカーネル関数の設計も進めていく。

(2) 木構造データに対する圧縮と知識発見

内包カーネル関数は導出を基にしているために木構造と密接な関係がある。本研究では木構造データに対するデータ圧縮や知識発見を通して、内包に関する知見を得ることを目的として、木構造データに対する圧縮と

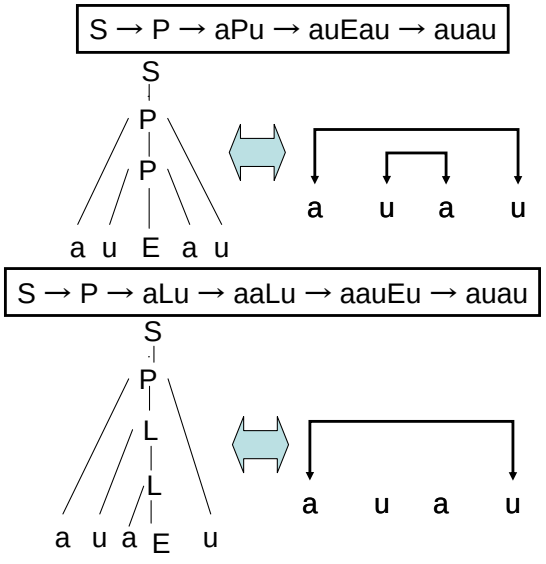
知識発見に関しても研究を行う。

４．研究成果

(1) 内包カーネル関数の設計

先行研究で行った非コードRNAを機能で分類する問題に対するカーネル関数の拡張、更なる解析を行った。下図は我々の設計した２次構造を利用したRNA配列に対するカーネル関数の設計である。文脈自由文法を利用してカーネル関数を設計し、更にこのカーネルを組合せることによって、様々な２次構造を学習できるようにしている。

$$S \rightarrow P \mid L \mid R \mid E$$
$$P \rightarrow xPy \mid xLy \mid xRy \mid xEy$$
$$L \rightarrow xP \mid xL \mid xR \mid xE$$
$$R \rightarrow Px \mid Lx \mid Rx \mid Ex$$
$$x \in \{a,u,c,g\}$$
$$(x,y) \in \{(a,u),(u,a),(c,g),(g,c)\}$$



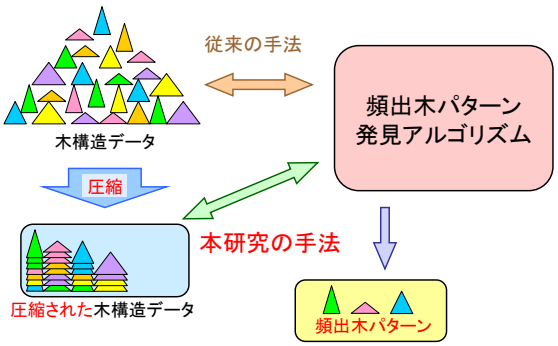
これ以外にも、RNA配列に対して、今までの我々の内包カーネル関数で考慮していなかったシュードノット構造を考慮するようにすべく修正を行ったが、パラメーラが多くなってしまい、現実のデータにはそれ程うまく適用することはできないと考えている。また、それ以外にも内包カーネル関数の設計を試みている。しかし、既存のカーネルと明確な差は見いだせていないため、更なる研究が必要である。

(2)圧縮を利用した頻出木パターン発見

半構造化文書をはじめとする木構造データに対する木文法圧縮を用いて木構造データからの頻出木パターンの発見をより効率化する手法を提案した。頻出木パターンの発見は大規模な木構造データから有用な情報を効率的に抽出することを目的としており，木構造データとしても表現できる HTML/XML 文書の普及とともに既に多くのアルゴリズムの研究が行われている．また，データ圧縮も大規模データに対して需要の高い技術であり，両者を組み合わせることは大規模な木構造データの情報活用のために意義のあることと考えられる．

多くの既存の発見アルゴリズムでは頻出木パターン候補の生成に対する工夫が行われてきた．本研究の提案手法はそれらの手法とは異なる新たなアプローチによる効率化の手法であり，既存の発見アルゴリズムの多くと競合することなく組み合わせることも可能である．さらに，そのような組み合わせによって既存のアルゴリズムの長所を生かしつつ相乗的な計算効率の改善も期待できる．本研究では根付き順序木と根付き無順序木に対して手法を提案している。

下図は我々の手法の模式図である。圧縮を利用して頻出木パターン発見の高速化を行っている。



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２件）

① Koichiro Doi and Akihiro Yamamoto, Kernel Functions Based on Derivation, New Frontiers in Applied Data Mining, Lecture Notes in Artificial Intelligence, 査読有, 5433, 2009, pp. 111-122

②Seiji Murakami, Koichiro Doi and Akihiro

Yamamoto, Proceedings of the 11th International Conference on Discovery Science (DS 2008), Lecture Notes in Artificial Intelligence, 査読有, 5255, 2008, pp.284-295

〔学会発表〕(計２件)

①土井晃一郎、圧縮された半構造化文書からの頻出木パターン発見、第２３回人工知能学会全国大会、2009 年 6 月 18 日、香川県高松市

② Koichiro Doi and Akihiro Yamamoto, Kernel Functions Based on Derivation, First International Workshop on Algorithms for Large-Scale Information Processing in Knowledge Discovery (ALSIP 2008), 2008 年 5 月 20 日, 大阪府大阪市

６．研究組織

(1)研究代表者

土井　晃一郎 (DOI KOICHIRO)

東京大学・大学院新領域創成科学研究科・特任講師

研究者番号：10345126

(2)研究分担者

なし

(3)連携研究者

なし