科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　２０　日現在

機関番号：　２４５０６
研究種目：　基盤研究(B)
研究期間：　2011～2013
課題番号：　２３３０００６１

研究課題名（和文）離散データ構造に対するカーネルの設計理論の構築

研究課題名（英文）A study on positive-definite and efficient kernels for structured data

研究代表者
申　吉浩（SHIN, Yoshihiro）

兵庫県立大学・応用情報科学研究科・教授

研究者番号：６０５２３５８７

交付決定額（研究期間全体）：（直接経費）　　15,000,000円、（間接経費）　　4,500,000円

研究成果の概要（和文）：ビッグデータなどの情報資産を将来の予測に役立てる機械学習技術の中で、カーネル法は多様なデータ構造に適用できる点で重要である。通常はベクター型のデータを仮定するが、現実には、DNA等の配列構造、蛋白質・構文木・XML文書等の木構造、各種ネットワーク等のグラフ構造など、構造をもつデータが多く存在する。本研究では、構造データを効率よく解析する手段として、構造カーネルの基礎理論構築と実用化に向けた研究を行った。具体的には、カーネルの必須要件である正定値性の判定理論、多様な構造への適用手法の開発、そして、研究者を対象者としたカーネル計算ユーティリティの構築とインターネット上での公開を行った。

研究成果の概要（英文）：Kernel method is an important field of machine learning research and allows us to leverage information assets like big data to make useful predictions in various applications. In addition to vector data,there exist huge amount of data that have structures. For example, DNA is an array of nucleotides; Protein, parse trees and XML documents are naturally structured as trees; Various kinds of networks are represented using the graph structure. In this regard, this project aims to establish a theory of kernels for structured data and practical techniques to apply kernels to structured data of the real applications. Specifically, we have developed a mathematical theory to investigate positive definiteness of kernels and various types of kernels that deal with a wide variety of structured data. Furthermore, we have developed a utility to compute kernels and have publicized it to researchers in the field of machine learning over the Internet.

研究分野：　情報学

科研費の分科・細目：　知能情報学

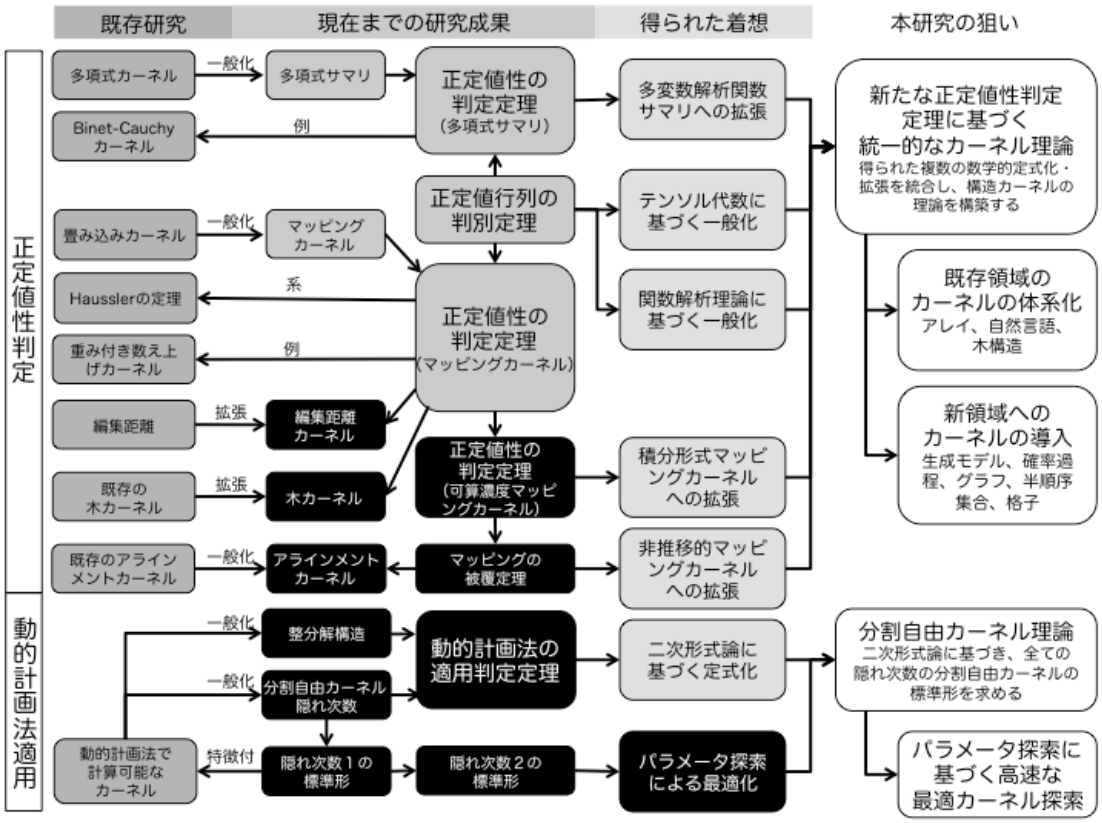
キーワード：　機械学習　構造化データ　カーネル法　動的計画法

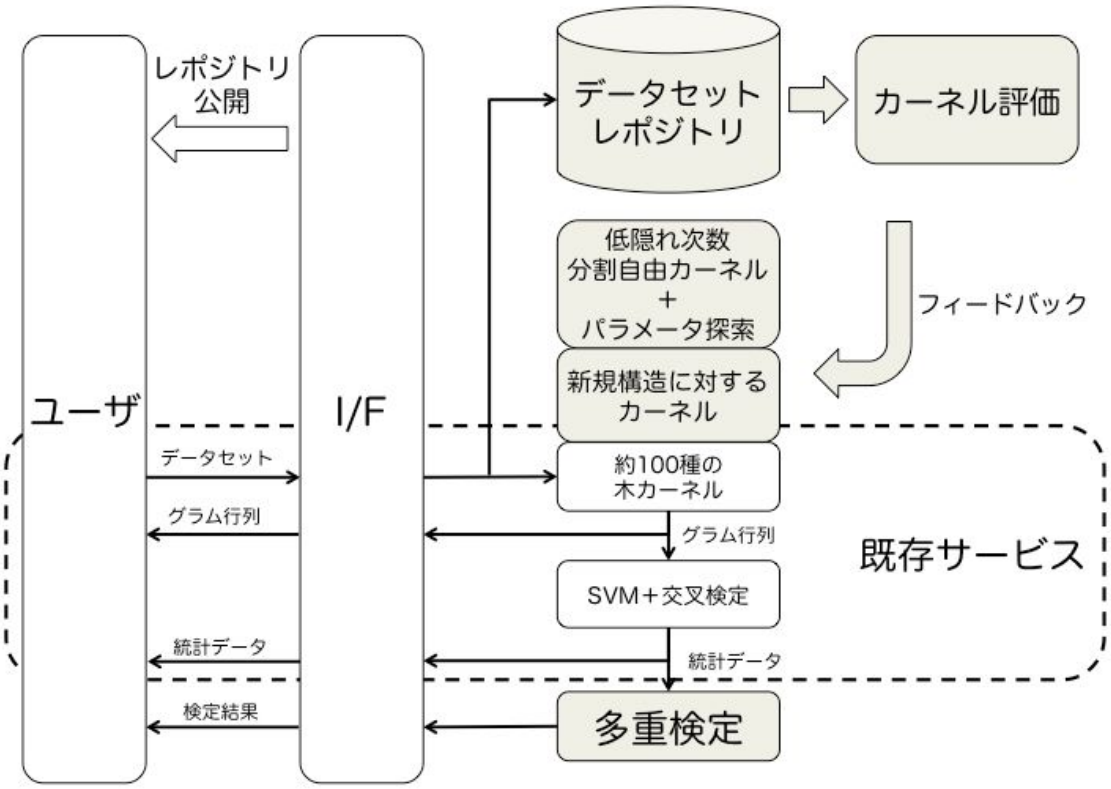
１．研究開始当初の背景
(1)SVM などの手法の開発により、多項式カーネル・ガウスカーネル等によるベクター型データに対するカーネル法の適用は一般化した一方、Haussler の畳み込みカーネルにより、主に配列型データに対する構造カーネルも徐々に利用されるようになって来た。
(2)一方、筆者等は、多項式カーネル及び畳み込みカーネルを顕著に一般化し、多項式サマリ・マッピングカーネルを提案し、構造カーネルの正定値性を判定するための一般的な理論の構築に取り組んだ。
(3)カーネルを利用するには、正定値性を証明すること、動的計画法等による効率的な計算手法を見つけることの二点が必須であるが、計算手法に対する理論的研究は遅れていた。

２．研究の目的
以下の 3 点を目的とした。
(1)構造カーネルの正定値性判定のための理論を発展させ、一般論と個別応用の両面で理論を構築すること。
(2)構造カーネルの効率的計算手法に関する理論を構築すること。
(3)研究領域において構造カーネルの利用の促進を目的として、構造カーネルの計算ユーティリティを開発・公開すること。

３．研究の方法
前記各目的に対して、以下の方法をとる。
(1)筆者等が既に得ている、多項式サマリ・マッピングカーネルの成果を更に発展させることで、理論の構築を行う。
(2)実用に供されているカーネルのほぼ全てが動的計画法により計算されている事実に鑑み、動的計画法の適用可能範囲を明らかにする方向で、理論の構築を行う。
(3)パラメータ設計に基づいて、多様な木カーネルの計算を網羅的・体系的に行うプログラムを開発し、計算ユーティリティとしてインターネット上で公開する。

４．研究成果
(1)正定値性判定、及び、計算可能性に関する理論の構築に関する研究成果の概略を、下図にまとめた。左欄の「既存研究」は、筆者等が研究を始める時点で文献等で知られていた、既存研究である。中央左欄の「現在までの研究成果」は、筆者等の研究成果を、本科研補助金による研究以前の成果（灰色の矩形で表示）と、本科研補助金による研究成果（黒色の矩形で表示）に分けて、示している。中央右欄の「得られた着想」は、本科研補助金による研究を通して得られた、将来の研究の方向を示す着想であり、筆者等がこれから直近に取り組む研究テーマとなる。右欄の「本研究の狙い」は、本科研補助金による研究成果と着想に基づく将来研究の狙いを示す。正定値性判定理論に関する研究成果は以下であり、一部は、機械学習におけるトップの国際会議である ICML 2011 に採択されている。
①既に得ている、多項式サマリ・マッピングカーネルの正定値性判定理論の整理を通して、これらの理論が、多変数解析理論・テンソル代数・関数解析理論などにより、顕著に抽象化・一般化できる可能性を見出し、一部理論化に着手した。この数学的な理論の一般化は、実用的には、具体的な構造に適用する構造カーネルの体系化、及び、新たな構造への応用の道を開くものである。
②従来のマッピングカーネルは、構造の有限性を仮定していた。実データを考えると、この仮定は妥当なものと思えたが、実は、加算無限濃度の構造を考えることにより、有限性を有する実データの解析にも役立つことが分かって来た。まず、現在は有限和により表現されているマッピングカーネルを積分形式に拡張することにより、確率分布等、生成モデルとして表現し得るデータに対して、生成モデルに即したカーネル定義が得られる可能性がある。更に、被覆定理、即ち、有限構造のデータに対して、加算無限濃度の構造の被覆を考えることにより、これまでのマッピングカーネルの正定値判定定理では証明できなかったクラスのカーネル（非推移的マッピングカーネル）の正定値性が証明できるようになる可能性がある。
③また、具体的な構造カーネルの開拓に関しては、今まで数種類しか知られていなかった（正定値性が証明されていなかった）木構造を対象としたカーネルに対し、一挙に百種類上の新規のカーネルの正定値性を示した。これらの新規カーネルは、実験を通して、その有効性が顕著であることが既に示されているものも含まれ、今後、有効性の検証を進めて行く。また、既に確立されている編集距離理論との関連を研究し、編集距離と特定のクラスのカーネルの間に密接な関係があることを明らかにした。このクラスのカーネルを、編集距離カーネルと呼ぶ。更に、編集距離理論では、文字列編集距離や一部の木編集距離に対してしか知られていなかったアラインメントの概念を一般化し、任意のグラフ構造に対するアラインメントの概念を導入し、更に、アラインメントに基づいて定義されるアラインメントカーネルを導入した。この研究の意義は、実は、編集コストの最小値として定義される編集距離に対して、カーネルは編集コストの分布そのものを評価するため、よりよい分析を行える可能性がある点に依拠している。この見通しは、実験結果からも、裏付けられている。
(2)動的計画法による構造カーネルの計算可能性に関する研究については、従来知られていた一部の畳み込みカーネルの計算可能性を一挙に拡大し、分割可能カーネルと整分解構造に基づいて定義される任意のマッピングカーネルが動的計画法によって計算でき





ることを示した。更に、従来動的計画法により計算可能であった畳み込みカーネルは、本研究で示された計算可能範囲の極く一部であり、実際、分割可能カーネルには、任意の自然数を値に取る不変量（隠れ次数）を定義でき、従来の計算可能カーネルは、隠れ次数１のケースに正確に対応する。一方、隠れ次数が２以上の分割可能カーネルに、有意義な

例が複数存在し、分割可能カーネルによる計算可能性の理論には、今後、カーネルの適用領域を拡大する基盤となる期待が持てる。更に、隠れ次数２の分割可能カーネルは、数個のパラメータで表現される標準形に帰着できることを示した。隠れ次数１の分割可能カーネルも一個のパラメータで表現されることから、この結果により、隠れ次数１及び２の分割自由カーネルを用いたマッピングカーネルの最適解を、パラメータ探索により体系的・網羅的に探索できることが分かる。この結果は、隠れ次数が３以上の場合にも敷衍できることが期待できるので、より広い範囲のカーネルに対して、効率的な最適カーネルの探索が可能となるものと考えられる。これらの結果は、機械学習のトップの国際会議である、ICML 2013 に採択され手いる。

(3)カーネル計算ユーティリティに関しては、総数 100 以上の木カーネルを実装したプログラムを開発した。このプログラムの開発には、本研究の成果である体系的なパラメータによる木カーネルの表現に基づいており、互いに独立となるように設計された３つのパラメータに値を設定することで、カーネル計算から交叉検定までを効率的に行うことができる。このプログラムは、インターネット上の計算ユーティリティとして、近々に、研究者を対象として公開する予定である（現在は、研究プロジェクト参加者に限定)。上図は、この計算ユーティリティの概要を示したもので、既存サービスと書かれている部分が、本科研補助金の研究成果として実現済みの部分である。灰色の矩形の部分は、今後の拡張を予定している部分となる。今後は、隠れ次数が２以上の分割自由カーネルに基づくカーネル、木以外の構造を取り扱うカーネルを追加することで、応用範囲の拡大を図る。また、利用者の許諾を得て、データセットを収集し、データセットレポシトリを構築する。データレポシトリはインターネットを介して利用者に公開する外、カーネルの性能評価を行うためのテストベッドとして利用し、ユーティリティで提供するカーネルの改善に役立てる。更に、複数のカーネルを比較するための、多重検定機能を導入し、利用者が適切なカーネルを選択する目的に供する。

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４件）

1. Alignment Kernels based on a Generalization of Alignments, K. Shin, IEICE Transactions on Information and Systems: Vol.E97-D, No.1, 1-10, 2014. 査読有. doi:10.1587/transinf.E97.D.1
2. A Theory of Subtree Matching and Tree Kernels based on the Edit Distance Concept, K. Shin, Machine Learning, Springer, To appear, 2014. 査読有. 採録決定済
3. Mapping kernels for infinite mapping systems, K. Shin, Journal of Machine Learning Research, Proceeding Track 20: 367–382, 2011. 査読有. http://jmlr.org/proceedings/papers/v20/shin11.html
4. 部分パスに基づいた木カーネル, 木村 大翼, 久保山 哲二, 渋谷 哲朗, 鹿島 久嗣, 人工知能学会論文誌: Vol.26, No.3, 473–482, 2011. 査読有. doi:10.1527/tjsai.26.473

〔学会発表〕（計１０件）

1. Edit distance and alignment kernels, K. Shin, In International Conference on Artificial Intelligence, Soft Computing (AISC 2013), 2013. 査読有, Feb. 18 - 20, Bangalore, India
2. A New Frontier of Kernel Design for Structured Data, K. Shin, In The 30th International Conference on Machine Learning (ICML 2013), ACM, To appear, 2013. 査読有, June 16 - 21, Atlanta, USA
3. Selecting tree kernels cleverly, K. Shin and T. Kuboyama, In Workshop on Data Discretization and Segmentation for Knowledge Discovery (DDS13), To appear, 2013. 査読有, October 27-28, 慶応大学(神奈川県横浜市)
4. Exploring social context from buz marketing site-commuinity mapping based on tree edit distance, S. Higuchi, T. Kuboyama, T. Hashimoto and K. Hirata , In PerCom Workshops 2013 : 187–192, 2013. 査読有. March 18 - 22, San Diego, USA
5. Dynamic labeling and tree kernels with gap penalties, K. Shin and T. Kuboyama, In The 6th International Conference on Soft Computing and Intelligent Systems, IEEE Explore, 2012. 査読有. November 20 - 24, 神戸国際会議場（兵庫県神戸市）
6. Mapping kernels for trees, K. Shin, M. Cuturi and T. Kuboyama, In The 28nd International Conference on Machine Learning (ICML 2011), ACM, 961 – 968, 2011. 査読有. June 28 - July 2, Belleview, Washington, USA
7. Partitionable Kernels for Mapping Kernels, K. Shin, In The 11th IEEE International Conference on Data Mining (ICDM 2011), 645–654, 2011. 査読有. December 11 - 14, Vancouver, Canada
8. Improved MAX SNP-Hard Results for Finding an Edit Distance between Unordered Trees, K. Hirata , Y. Yamamoto and T. Kuboyama, In CPM 2011, 402–415 2011. 査読有. June 27 - 29, Palermo, Italy
9. On Computing Tractable Varlations of Unordered Tree Edit Distance with Network Algorithms, Y. Yamamoto, K. Hirata and T.

Kuboyama, In JSAI-isAI Workshops 2011: 211–223 2011. 査読有. December 1 - 2, 高松サンポートホール（香川県高松市）
10. A Subpath Kernel for Rooted Unrooted Trees, D. Kimura, T. Kuboyama, T. Shibuya and H. Kashima, In PAKDD (1) 2011: 62–74 2011. 査読有. May 24 - 27, Shenzhen, China

〔その他〕
ホームページ等
http://www.kamuy-nitay.com/treekernel

６．研究組織
(1)研究代表者
申　吉浩（SHIN, Yoshihiro）
兵庫県立大学・応用情報科学研究科・教授
研究者番号：60523587

(2)研究分担者
岡本　洋（OKAMOTO, Hiroshi）
独立行政法人理化学研究所・脳回路機能理論研究チーム・研究員
研究者番号： 00374067

有村　博紀(ARIMURA, Hiroki)
北海道大学・情報科学研究科・教授
研究者番号：20222763

坂本　比呂志(SAKAMOTO, Hiroshi)
九州工業大学・情報工学研究院・教授
研究者番号：50315123

久保山　哲二(KUBOYAMA, Tetsuji)
学習院大学・計算機センター・教授
研究者番号：80302660

Cuturi Marco(CUTURI, Marco)
京都大学・情報学研究科・准教授
研究者番号：80597344