様式 C-19

科学研究費補助金研究成果報告書

平成 23 年 6 月 9 日現在

機関番号：32606
研究種目：基盤研究(C)
研究期間：2008～2010
課題番号：20500126
研究課題名（和文）　離散構造間の類似度設計と機械学習への応用

研究課題名（英文）　Designing similarity measures for discrete data structures, and their applications to machine learning

研究代表者
久保山 哲二（KUBOYAMA TETSUJI）
学習院大学 計算機センター・准教授
研究者番号：80302660

研究成果の概要（和文）：本研究の目的は、文字列や木構造などの非数値データ間の距離や類似度を測るための一般的な手法を開発することにある。なかでも、類以度がカーネル関数と呼ばれるクラスである場合には、従来、数値データを対象としていた統計手法や機械学習手法を、カーネル法を用いて非数値データにも適用することができるようになる。本研究では、離散データ構造の類似度設計フレームワークとして広く使われている畳込みカーネルを、より実際的かつ一般的な形に拡張し、構造間の対応関係の数え上げによってカーネル関数を設計するマッピングカーネルを提案した。また、木構造間の共通部分構造を数え上げる手法を用いて木構造を対象とした様々な類以度(カーネル関数)を提案した。

研究成果の概要（英文）：The similarity measures called kernel functions for discrete data structures such as strings, trees, and graphs allow for statistical analysis and machine learning for non-numerical variables. In this study, a novel and general framework for designing kernel functions for discrete data structures has been developed. The framework is a generalization of Haussler's convolution kernel by counting all possible structural correspondences between two structures. Moreover, a diversity of similarity measures for trees have been developed based on counting their common substructures.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	1,300,000	390,000	1,690,000
2009 年度	1,100,000	330,000	1,430,000
2010 年度	1,100,000	330,000	1,430,000
総　計	3,500,000	1,050,000	4,550,000

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：機械学習，類以度，木構造，カーネル法

１．研究開始当初の背景

(1) 離散構造間の近似パターン照合

生物情報学や自然言語解析、画像解析をはじめとするさまざまな分野で知識発見や高速検索のための要素技術として、文字列や木構造等の離散データ構造間の近似パターン照合や類似度の計算手法が開発されてきた。なかでも、編集距離の概念を用いた手法は、パターン照合や類似度計算のための標準的な手法として広く受け入れられており、様々な応用分野で実際に使われている。

本研究代表者と研究協力者の申吉浩教授(兵庫県立大学)は、木の編集距離の意味論に着目し、従来、操作的に定式化されてきた木の編集距離を、順序代数を用いた順序集合間の写像の性質として厳密に定式化した。これ

により、既存研究に含まれていた様々な誤謬や問題点を解消し、従来別々の意味を持つと思われていた複数の木の編集距離尺度やその計算アルゴリズムが実は等価であることや、既存の複数の木の編集距離尺度には構造比較の感度(近似の度合い) に応じたきれいなクラス階層が存在すること、計算量と構造比較の感度には密接なかかわりがあることなどを明らかにした。

(2) カーネル関数の設計フレームワーク

近年のサポートベクターマシン(SVM) に代表される学習器を用いたカーネル法による機械学習の爆発的な流行にともない、さまざまな種類のデータ間の類似度(カーネル関数) が提案されている。直観に基づき類似度を設計しても、カーネル関数としての要件は必ずしも満たされないため、カーネル関数の設計に際しては、いくつかの設計の指針となるフレームワークが提案されている。離散データ構造については、強力な設計フレームワークとして畳み込みカーネルが知られており、広くカーネル関数の設計に用いられている。しかし、畳み込みカーネルは「離散構造全体の類似度を部分構造の類似度の総和によって計算する」という必ずしも本来の定義に一致しない概念的な理解が流布しているために、畳み込みカーネルに基づいて設計したと謳って提案された類似度が、畳み込みカーネルの定義から外れているケースも散見される。本研究代表者らは、既存の離散データ構造を対象としたいくつかのカーネル関数が畳み込みカーネルの範疇を超えていることを示している。

２．研究の目的

研究の目的は、次に示す「離散データ構造の近似パターン照合」と、「カーネル関数の強力な設計フレームワーク」の2 つに大別される。

(1) 離散構造間の近似パターン照合

離散構造間の編集距離の意味論の構築と、高速なアルゴリズムの開発を目的とする。具体的には以下の2 つの問題の解決を目指す。

① 動的計画法を用いない木の編集距離アルゴリズムの開発

動的計画法により木構造を再帰的に分解してゆくタイプの木の編集距離アルゴリズムの時間計算量は、木のノード数nに対して$\Theta(n^3)$であることがわかっている。本研究では動的計画法によらないより高速な木の編集距離の近似アルゴリズムの開発を行う。具体的には、木構造を部分木に分解し、部分木の頻度ベクトル間の距離により編集距離を近似する。

また、無順序木の編集距離やカーネル関数については、全ての部分構造をチェックする方法については計算困難性が証明されているため、新たな観点から多項式時間のアルゴリズムを開発する。

(2) カーネル関数の設計フレームワーク

既存のカーネル設計のフレームワークとして知られている多項式カーネルや畳み込みカーネルを、理論的に拡張し、より一般的なカーネル設計のフレームワークを提案する。また、木構造よりも表現力の高い離散構造に対する距離・類似度を設計し、これらを計算するためのアルゴリズムを開発する。

３．研究の方法

研究の 2 つの柱である「離散データ構造間の近似パターン照合」と「カーネル関数の強力な設計フレームワーク」それぞれについて、基礎理論の構築とアルゴリズム開発の両輪で研究を進めた。「離散データ構造間の近似パターン照合」ついては、部分構造の数え上げによる類以度の設計手法について研究協力者の平田耕一准教授(九州工業大学)と共に研究を進めた。

また、「カーネル関数の強力な設計フレームワーク」については、木構造を対象としたカーネル関数の設計で培ってきたカーネル関数設計のフレームワークを、一般の離散データ構造に拡張し、開発した手法の有効性を実証するために、DNA 配列、糖鎖構造等の実データに対して提案手法を適用した。なお、本研究は研究協力者の申吉浩教授((研究開始時)カーネギーメロン大学、(現)兵庫県立大学)と 2 人で遂行した。

また、応用にあたっては、バイオインフォマティックスやパターンマッチアルゴリズムの分野で実績のある国内外の研究者らと広く情報交換することにより、各々の研究領域に対する理解を深め、各分野の進んだ研究結果を積極的に取り入れながら研究を進めた。

４．研究成果

① 木構造を対象とした類似度・カーネル関数の開発

木構造の中でも子ノード間に順序のない、いわゆる無順序木を対象にカーネル関数を設計した。無順序間の一般的な木カーネルについては、2007 年に鹿島・坂本・小柳によって、その計算困難性が指摘されていた。本研究では、二つの木構造間に共通に含まれる二分木構造のみを効率よく数え上げることによって、共通構造の組合せ爆発の問題に対処し、カーネル関数を計算する高速なアルゴリズムを提案した。

順序木については、二つの木構造間の標準的な編集距離アルゴリズムとして、木構造の大きさの 3 乗オーダーの計算量をもつアルゴリズムが知られている。動的計画法を用いて木構造の対応関係を数え上げるカーネル関数にも、同様にこの計算量の壁があるため、異なる類似度計算のアプローチによる近似計算で高速化を行った。

具体的には、木構造の様々な特徴を数え上げることで、様々な木の編集距離の近似尺度やカーネル関数を提案し、糖鎖データや人工的に生成した木構造を対象に実験を行い、各々の距離と類似度の特徴を明らかにした。

② 離散データ構造のための類似度(カーネル関数)設計フレームワークであるマッピングカーネルの開発

木構造に対して開発した新しいカーネル関数の設計手法であるマッピングカーネルのフレームワークを、兵庫県立大学の申吉浩教授との共同研究により木構造に限定しない一般的な形に拡張した。これにより、離散データ構造のカーネル設計として現在もっとも一般的な Haussler の畳み込みカーネルの設計条件を緩和し、より一般的なフレームワークであるマッピングカーネルを開発した。

2 つの離散データ構造 x, y について、畳み込みカーネルでは、まず、各々の構造を部分構造に分解し、次に、部分構造に対して定義されているカーネル関数を用いて、全体構造に対するカーネル値を計算する。各々の部分構造の集合をそれぞれ $S(x)$, $S(y)$ とおくと、畳み込みカーネルでは積集合 $S(x)\times S(y)$ 上でカーネル関数の値が定義されていなくてはならない。構造 1 次の畳み込みカーネルは、部分構造間のカーネル関数を $\mathbf{K}'(x', y')$ と置くと次式で示される。

$$\mathbf{K}(x, y) = \sum_{(x', y') \in S(x)\times S(y)} \mathbf{K}'(x', y')$$

これに対して、マッピングカーネルでは、次式のように $S(x)\times S(y)$ の部分集合 M 上でカーネル関数が定義されていればよい。

$$\mathbf{K}(x, y) = \sum_{(x', y') \in M \subseteq S(x)\times S(y)} \mathbf{K}'(x', y')$$

一見わずかな違いであるが、類似度の設計の観点からは畳み込みカーネルの大幅な拡張になっており、一部の部分構造間のカーネル値が未定義である場合にもカーネル関数の設計ができ、柔軟に共通構造を定義し類似度を設計することが可能になる。ただし、そのままではカーネル関数としての半正定値性の条件をみたさず、$S(x)\times S(y)$ の部分集合が推移性を満たすことが必要十分条件となる。

畳込みカーネルが各構造の部分集合の数え上げに基づく方法であると捉えると、マッピングカーネルは二つの構造間の対応関係の数え上げによる方法である。従来より離散データ構造のカーネル計算に用いられてきた動的計画法による実装は、対応関係の数え上げの観点から素直に解釈でき、今まで畳込みカーネルの枠組みでは解釈が煩雑になっていたカーネル関数もマッピングカーネルの枠組みで素直に解釈できるようになった。

木構造を超える離散データ構造に対する距離・類似度の設計については、マッピングカーネルにより一般的なフレームワークは示せた。ただし、実際に木構造を超えた一般的な構造への実装を示すには至っておらず、この点については今後の課題である。

これらの成果は、機械学習で権威のある国際会議 ICML で発表された。以来、すでに海外の研究において実際にカーネル関数の設計やカーネル関数の半正定値性の証明に用いられていることを確認している。

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 17 件）

① Sata, K., Hirata, K., Ito, K. & Kuboyama, T.: Discovering Global Propagation for Influenza A Viruses Based on Robinson-Foulds Distance between Phylogenetic Trees, Proc. Annual International Conference on BioInformatics and Computational Biology (BICB), 2011, pp. B12-B17.

② Otsuka, Y., Miyahara, T. & Kuboyama, T.: Learnig of Multiple Tree Structured Patterns Using Clustering and Evolution, Proc. IADIS International Conference Information Systems 2011, pp. 227-231.

③ Yamamoto, Y., Hirata, K. & Kuboyama, T.:

A Bottom-Up Edit Distance between Rooted Labeled Trees, Proc. 7th Workshop on Learning with Logics and Logics for Learning (LLLL), 2011, pp. 26-33.

④ Kuboyama, T. & Hirata, K.: Broom Distance between Rooted Labeled Trees, Proc. 7th Workshop on Learning with Logics and Logics for Learning (LLLL), 2011, pp. 34-41.

⑤ Yoshida, K., Miyahara, T. & Kuboyama, T.: Evolution of Multiple Tree Structured Patterns using Soft Clustering, Proc. 2nd International Conference on Computer and Automation Engineering (ICCAE), 2010, Vol. 5, pp. 749-753.

⑥ Aratsu, T., Hirata, K. & Kuboyama, T.: Approximating Tree Edit Distance through String Edit Distance for Binary Tree Codes, Fundamenta Informaticae, 2010, Vol. 101(3), pp. 157-171.

⑦ Shin, K. & Kuboyama, T.: A Generalization of Haussler's Convolution Kernel - Mapping Kernel and Its Application to Tree Kernels, Journal of Computer Science and Technology, 2010, Vol. 25(5), pp. 1040-1054.

⑧ Horiike, T., Takahashi, Y., Kuboyama, T. & Sakamoto, H.: Extracting Research Communities by Improved Maximum Flow Algorithm, Knowledge-Based and Intelligent Information and Engineering Systems, 13th International Conference (KES), 2009, LNCS 5712, pp. 472-479.

⑨ Aratsu, T., Hirata, K. & Kuboyama, T.: Approximating Tree Edit Distance through String Edit Distance for Binary Tree Codes, SOFSEM 2009: Theory and Practice of Computer Science, 35th Conference on Current Trends in Theory and Practice of Computer Science, 2009, LNCS 5404, pp. 93-104.

⑩ Sata, K., Hirata, K., Ito, K. & Kuboyama, T.: Discovering Networks for Global Propagation of Influenza A (H3N2) Viruses by Clustering, Knowledge-Based and Intelligent Information and Engineering Systems, 13th International Conference (KES), 2009, LNCS 5712, pp. 490-497.

⑪ Shin, K. & Kuboyama, T.: Polynomial summaries of positive semidefinite kernels, Theoretical Compututer Science, 2009, Vol. 410(19), pp. 1847-1862.

⑫ 申 吉浩 & 久保山 哲二: Haussler畳み込みカーネルの一般化と応用 : マッピングカーネル, 人工知能学会論文誌, 人工知能学会, 2009, Vol. 24(2), pp. 263-271.

⑬ Kuboyama, T., Hirata, K. & Aoki-Kinoshita, K.F.: An Efficient Unordered Tree Kernel and Its Application to Glycan Classification, Proc. 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD), 2008, LNCS 5012, pp. 184-195.

⑭ Aratsu, T., Hirata, K. & Kuboyama, T.: Sibling Distance for Rooted Labeled Trees, New Frontiers in Applied Data Mining, PAKDD 2008 International Workshops, 2008, LNCS 5433, pp. 99-110.

⑮ Nagamine, M., Miyahara, T., Kuboyama, T., Ueda, H. & Takahashi, K.: Evolution of Multiple Tree Structured Patterns from Tree-Structured Data Using Clustering, AI 2008: Advances in Artificial Intelligence, 21st Australasian Joint Conference on Artificial Intelligence, 2008, LNCS 5360, pp. 500-511

⑯ Shin, K. & Kuboyama, T.: Kernels Based on Distributions of Agreement Subtrees, AI 2008: Advances in Artificial Intelligence, 21st Australasian Joint Conference on Artificial Intelligence, 2008, LNCS 5360, pp. 236-246.

⑰ Shin, K. & Kuboyama, T.: A generalization of Haussler's convolution kernel: mapping kernel, Machine Learning, Proc. of the 25th International Conference (ICML), 2008, pp. 944-951.

〔学会発表〕（計 9 件）

① Kuboyama, T., Ito, K., Hirata, K. & Sakamoto, H.: Predicting Mutations of Influenza Virus HA Genes Using Dimensionality Reduction of Hamming Distance Space, Proc. Annual International Conference on BioInformatics and Computational Biology (BICB), 2011-02-28, pp. B43, シンガポール..

② 平田耕一, 山本恭之, 久保山哲二: On MAX SNP-hard results for unordered tree edit distance, 第 80 回人工知能基本問題研究会, 2010-11-17, pp. 33-38, 東京.

③ 申吉浩, 久保山哲二: 自由度 1 及び 2 の分割自由カーネル, 第 79 回人工知能基本問題研究会, 2010-09-25, Vol. 79, pp. 35-39, 北海道大学.

④ 久保山哲二, 申吉浩, 伊藤公人: アミノ酸の高次元符号化によるインフルエンザウイルスの抗原変異予測, 第 78 回人工知能基本問題研究会, 2010-07-31, pp. 1-7, 兵庫県立大学.

⑤ 久保山哲二, 伊藤公人: ハミング距離空間の次元削減によるインフルエンザウイルス遺伝子変異の解析, 第 77 回人工知能基本問題研究会, 2010-03-17, pp. 91-95,

北海道大学.
⑥ 吉田健吾，宮原哲浩，久保山哲二: Evolution of multiple tree structured patterns using soft clustering, 第77回人工知能基本問題研究会, 2010-03-17, pp. 49-54, 北海道大学.
⑦ 申吉浩, 久保山哲二: 第 73 回木カーネルの構成のためのフレームワークとサーベイ，人工知能基本問題研究会, 2009-03-14, pp. 41-48, 学習院大学.
⑧ 長嶺将俊，宮原哲浩，久保山哲二: Evolution of multiple tree structured patterns using clustering, 第 73 回人工知能基本問題研究会, 2009-03-13, pp. 35-40, 学習院大学
⑨ 荒津拓, 平田耕一, 久保山哲二: 二分木符号の文字列編集距離による木の編集距離の近似, 第 71 回人工知能基本問題研究会, 2008-09-17, pp. 57-62, 北海道大学.

６．研究組織

(1)研究代表者

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