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

A Generalization of Haussler's Convolution Kernel Mapping Kernel

申 吉浩
Kilho Shin

カーネギーメロン大学日本校
Carnegie Mellon CyLab Japan
yshin@cmuj.jp

久保山 哲二
Tetsuji Kuboyama

学習院大学
Gakushuin University
kuboyama-jsaih@tk.cc.gakushuin.ac.jp

keywords: machine learning, kernel method, convolution kernel

Summary

Haussler's convolution kernel provides a successful framework for engineering new positive semidefinite kernels, and has been applied to a wide range of data types and applications. In the framework, each data object represents a finite set of finer grained components. Then, Haussler's convolution kernel takes a pair of data objects as input, and returns the sum of the return values of the predetermined primitive positive semidefinite kernel calculated for all the possible pairs of the components of the input data objects. On the other hand, the *mapping kernel* that we introduce in this paper is a natural generalization of Haussler's convolution kernel, in that the input to the primitive kernel moves over a predetermined subset rather than the entire cross product. Although we have plural instances of the mapping kernel in the literature, their positive semidefiniteness was investigated in case-by-case manners, and worse yet, was sometimes incorrectly concluded. In fact, there exists a simple and easily checkable necessary and sufficient condition, which is generic in the sense that it enables us to investigate the positive semidefiniteness of an arbitrary instance of the mapping kernel. This paper presents and proves the validity of the condition. In addition, we introduce two important instances of the mapping kernel, which we refer to as the *size-of-index-structure-distribution* kernel and the *edit-cost-distribution* kernel. Both of them are naturally derived from well known (dis)similarity measurements in the literature (the maximum agreement tree, the edit distance), and are reasonably expected to improve the performance of the existing measures by evaluating their distributional features rather than their peak (maximum/minimum) features.

1. は じ め に

Haussler による畳み込みカーネル [Haussler 99] は、普遍的カーネル関数を個別のアプリケーションにカスタマイズするための一般的なフレームワークとして、広い応用範囲をもつことが知られている。本稿では、Haussler の畳み込みカーネルを式(1) で定義される関数であるとする。この定義は、原論文 [Haussler 99] での定義と同値であるが(2.2 節)、原論文の定義に比べてよりよい見通しを与える。 χ をデータポイントの集合、 χ' をカーネル関数 $k: \chi' \times \chi' \rightarrow \mathbb{R}$ が定義される空間とし、各データポイント $x \in \chi$ に対して、有限集合 $\chi'_x \subseteq \chi'$ が与えられているとする。 χ' を各データポイントを構造データと見た時の部分構造の全体集合、 χ'_x を x を構成する部分構造の集合と考えると分かりやすいかも知れない [鹿島 06]。Haussler の畳み込みカーネル $K: \chi \times \chi \rightarrow \mathbb{R}$ を以下のよ

うに定義する。

$$K(x, y) = \sum_{(x', y') \in \chi'_x \times \chi'_y} k(x', y') \quad (1)$$

Haussler の業績は、 $k(x', y')$ が半正定値であるならば、 $K(x, y)$ も半正定値であることを示した点にあり、この結果を利用して多くの半正定値カーネルが Haussler の畳み込みカーネルとして実現されてきた [Lodhi 02, Collins 01, Suzuki 04]。

一方、本稿で提案するマッピングカーネルは、Haussler の畳み込みカーネルのごく自然な拡張としてのように定義される。 $M_{x, y} \subseteq \chi'_x \times \chi'_y$ はデータポイントのペア (x, y) 毎に定められる有限集合であり、 $M_{x, y}$ が χ'_x から χ'_y への多値写像を定めることから、式(2) のカーネルをマッピングカーネルと呼ぶ。

$$K(x, y) = \sum_{(x', y') \in M_{x, y}} k(x', y') \quad (2)$$

本稿の第一の貢献は、マッピングカーネルが半正定値となるための条件を提示する点にある。即ち、任意に選んだ半正定値 $k(x', y')$ に対して式(2)による $K(x, y)$ が常に半正定値となる必要十分条件は、 $\{M_{x,y} \mid (x, y) \in \chi^2\}$ が推移的であること、即ち、 $(x', y') \in M_{x,y} \wedge (y', z') \in M_{y,z} \Rightarrow (x', z') \in M_{x,z}$ が成立することである。当然のことながら、Haussler の畳み込みカーネルは、 $\{M_{x,y} = \chi'_x \times \chi'_y\}$ とした時のマッピングカーネルと一致し、Haussler の結果は本稿の結果の特殊な例となる。

個別の例としては、マッピングカーネルは過去の文献にも見られる。

- 鹿島等 [Kashima 02, 鹿島 06] は、木カーネルの一種である *elastic tree kernel* を提案し、その半正定値性は Haussler の定理によるとしていた。実際には、*elastic tree kernel* はマッピングカーネルの一例であり、その半正定値性は本稿の結果から導かれるべきである。*elastic tree kernel* におけるマッピングシステムは推移的であり、*elastic tree kernel* が半正定値であるという結果は正しいものであった。
- Zien 等は、DNS 配列中の翻訳開始点を同定することを目的として、*codon-improved kernel* [Zien 00] を導入した。Zien 等は、*codon-improved kernel* を多項式カーネルの一例であると考えていたようであり、そのため、常に半正定値性が成立するものとしていた。実際には、特定の条件のもとで、*codon-improved kernel* はマッピングカーネルと解釈され、その時のマッピングシステムが推移的でないため、実際に半正定値ではないことが確かめられる。

Elastic tree kernel と *codon-improved kernel* とは、誤った根拠に基づいて半正定値性が判定されていたカーネルの例であり、本稿の結果を用いることにより、正しく半正定値性を判定することができる。

一方、Menchetti 等 [Menchetti 05] による *weighted decomposition kernel* や久保山等 [Kuboyama 06] による *tree mapping kernel* も、マッピングカーネルの例となるカーネル関数であるが、その半正定値性は、畳み込みカーネルとの違いを意識して、正しく証明されている。但し、証明はあくまで個別であり一般性はなかった。

本稿では、カーネル関数の一般的なクラスとしてマッピングカーネルを定義し、マッピングカーネルが半正定値となるための必要十分条件を示す。2 章において、マッピングカーネルの形式的な定義を与え、本稿の主定理を述べる。主定理の証明は 4 章で与える。主定理で述べる条件は直感的に理解しやすく、また、検証も容易である。その観点から、新たなカーネルを設計する際に有用な道具となるものと考えられる。

3 章ではマッピングカーネルの実例を示す。具体的には、従来からよく知られている類似度の尺度の実例 (編集距離と最大共通構造) から、マッピングカーネルを構成するための手法を示す。

編集距離は、一方の構造を他方に変換する際の編集コストの最小値として定義され、類似度の尺度として広い応用範囲をもつ。また、二つの構造の間の共通構造の大きさの最大値は、構造間の類似性を測る尺度として利用できることが知られている。*Maximum agreement subtree* はその代表的な例である。編集距離や最大共通構造では、構造のペア毎に分布が定まる指標 (編集コストや共通構造の大きさ) の最小値或いは最大値をもって定義されるので、指標の分布がもつ情報が利用されずに捨てられてしまう嫌がある。一方、本稿で示す手法では、これら指標の分布が持つ情報を構造の類似性の判定に活用するので、より有効な類似度関数を与えるものと期待できる。5 章では、鹿島等 [Kashima 02, 鹿島 06] による *elastic tree kernel* との比較を、糖鎖データセットを用いた実験により示す。

これらの分布カーネルの半正定値性の判定には本稿の主定理を利用するが、判定結果は指標のクラスに依存する。3.1 節及び 3.2 節では、*Tai* 編集距離 (*Tai edit distance*) 及び *maximum agreement subtree* に基づく分布カーネルは常に半正定値になるのに対し、*less-constrained tree edit distance* [Lu 01] 及び *maximum refinement trees* [Hein 96] に基づく分布カーネルに対しては、 $\{M_{x,y}\}$ が推移的にはならないことを見る。

2. マッピングカーネル

この節では、2.1 節で半正定値カーネル、2.2 節で Haussler 畳み込みカーネルについて概観した後、2.3 節においてマッピングカーネルの形式的定義を与える。

2.1 半正定値カーネル

対称なカーネル関数 $K: \chi \times \chi \rightarrow \mathbb{R}$ が半正定値であるとは、任意のデータポイント $x_1, \dots, x_n \in \chi$ に対して、Gram 行列 $G = [K(x_i, x_j)]_{i,j=1,\dots,n}$ が半正定値であることと定義される (定義 1)。カーネル関数の半正定値性は、再生カーネルヒルベルト空間 (reproducing kernel Hilbert space) が存在するための条件であり、特に、 χ が有限集合であるとする、写像 $\Phi: \chi \rightarrow \mathbb{R}^N$ が存在して、 $K(x, y) = \Phi(x)\Phi(y)^T$ が成り立つこと、即ち、 $K(x, y)$ があるユークリッド空間内での内積を与えていることと同値である。

定義 1. 実正方形行列 A が対称行列であり (i.e. $A^T = A$)、かつ、互いに同値な下記の条件を満足する時、 A は半正定値であるという。但し、 $\dim A = n$ とする。

- (1) 任意の $\forall (c_1, \dots, c_n) \in \mathbb{R}^n$ に対して、 $(c_1, \dots, c_n)A(c_1, \dots, c_n)^T \geq 0$ が成り立つ。
- (2) A の固有値は全て非負である。
- (3) n 次元の正規行列 $P (P^T P = E_n)$ が存在して、 $P^T A P$ が非負成分のみを含む対角行列となる。
- (4) m 行 n 列の実行列 B が存在して、 $A = B^T B$ が成

り立つ．

- (5) 任意の $\alpha_1, \dots, \alpha_m \in \{1, \dots, n\}$ に対して， $\det A[\alpha_1, \dots, \alpha_m] \geq 0$ が成り立つ．但し， $A[\alpha_1, \dots, \alpha_m]$ は (i, j) 成分が $A_{\alpha_i \alpha_j}$ と等しい m 次元正方行列とする．

2.2 Haussler 畳み込みカーネル

系 1 は R -畳み込みカーネルに関する Haussler の定理 [Haussler 99, Theorem 1] である．一方，定理 1 は Haussler の定理の $D = 1$ の場合であるが，興味深いことに系 1 は定理 1 から導かれる．

定理 1. カーネル関数 $k: \chi' \times \chi' \rightarrow \mathbb{R}$ は半正定値であるとする．関係 $R \subseteq \chi' \times \chi$ に関して，式(3)により定義されるカーネル関数 $K: \chi \times \chi \rightarrow \mathbb{R}$ もまた半正定値である．

$$K(x, y) = \sum_{(x', x) \in R} \sum_{(y', y) \in R} k(x', y') \quad (3)$$

系 1 ([Haussler 99]). 添字 $d = 1, \dots, D$ に対して， $k_d: \chi'_d \times \chi'_d \rightarrow \mathbb{R}$ を半正定値カーネル関数とする．

関係 $R \subset \chi'_1 \times \dots \times \chi'_D \times \chi$ に関して，下式によって定義されるカーネル関数 $K: \chi \times \chi \rightarrow \mathbb{R}$ もまた，半正定値である．

$$K(x, y) = \sum_{(x'_1, \dots, x'_D, x) \in R} \sum_{(y'_1, \dots, y'_D, y) \in R} \prod_{d=1}^D k_d(x'_d, y'_d)$$

証明 $\chi' = \chi'_1 \times \dots \times \chi'_D$ とする．系の主張は， $\prod_{d=1}^D k_d(x'_d, y'_d)$ は半正定値である事実 ([Cristianini 00] 等) と，定理 1 から直ちに導かれる． ■

2.3 マッピングカーネルの定義と主定理

$\chi'_x = \{x' \in \chi' \mid (x', x) \in R\}$ とする時，式(3)は式(1)と書き換えられる．一方，式(2)で与えられるマッピングカーネルは，式(1)において， (x', y') が $\chi'_x \times \chi'_y$ の部分集合である $M_{x, y}$ 上を動くとした時に定義される関数である．定理 2 では，任意の半正定値原始カーネル k に対してマッピングカーネルが半正定値となるための必要十分条件が，マッピングシステム $\{M_{x, y} \mid x, y \in \chi\}$ が推移的(定義 3)であることを主張する．

註 1. $\{M_{x, y}\}$ が推移的でなければ，マッピングカーネル $K(x, y)$ が半正定値とはならないような半正定値原始カーネル $k(x', y')$ が存在する (4.4 節)．

註 2. また，推移的でない $\{M_{x, y}\}$ と半正定値原始カーネル $k(x', y')$ に対して，マッピングカーネル $K(x, y)$ が半正定値となる例も存在するが，これは定理 2 の主張に矛盾するものではない．このような例として，Leslie 等 [Leslie 04] による (k, m) -mismatch kernel $K_{(k, m)}(x, y)$ は半正定値であるが， $M_{x, y}$ を文字列 x 及び y 中の k -mer のペア (x', y') の集合で， x' と y' の Hamming 距離が $2m$ 以下

となるように定めると，

$$K_{(k, m)}(x, y) = \sum_{(x', y') \in M_{x, y}} K_{(k, m)}(x', y')$$

とマッピングカーネルとして定義できるが， $\{M_{x, y}\}$ が推移的でないのは明らかである．

以下で，マッピングカーネルを形式的に定義する．

定義 2. マッピングシステム $\mathcal{M}$ を，三つ組 $(\chi, \{\chi'_x \mid x \in \chi\}, \{M_{x, y} \subseteq \chi'_x \times \chi'_y \mid (x, y) \in \chi^2\})$ として定義する．但し， $(x', y') \in M_{x, y}$ であるならば $(y', x') \in M_{y, x}$ であり，かつ， $|M_{x, y}| < \infty$ が成り立つ．

定義 3. マッピングシステム $(\chi, \{\chi'_x\}, \{M_{x, y}\})$ が推移的であるとは，任意の $x_i \in \chi$ and $x'_i \in \chi'_{x_i}$ ($i = 1, 2, 3$) に対して以下が成り立つことと定義する．

$$(x'_1, x'_2) \in M_{x_1, x_2} \wedge (x'_2, x'_3) \in M_{x_2, x_3} \Rightarrow (x'_1, x'_3) \in M_{x_1, x_3}$$

定義 4. $k: \chi' \times \chi' \rightarrow \mathbb{R}$ を任意の半正定値カーネル(原始カーネル)， $\gamma_x: \chi'_x \rightarrow \chi'$ を任意の写像(射影)であるとする時，マッピングシステム $(\chi, \{\chi'_x\}, \{M_{x, y}\})$ に関する評価システム $\mathcal{E}$ を，三つ組 $(\chi', k, \{\gamma_x \mid x \in \chi\})$ と定義する．

定義 5. マッピングシステム $\mathcal{M} = (\chi, \{\chi'_x\}, \{M_{x, y}\})$ と評価システム $\mathcal{E} = (\chi', k, \{\gamma_x\})$ から導出されるマッピングカーネルを，式(4)により定義する．

$$K(x, y) = \sum_{(x', y') \in M_{x, y}} k(\gamma_x(x'), \gamma_y(y')) \quad (4)$$

評価システムをマッピングシステムと独立に定義する理由は，半正定値原始カーネルを多様な構造に適用する為である．例えば，post order traversal でラベル付部分木を文字列に写像するように γ_x を定めることで，文字列間の一致関数を原始カーネルとして利用できる．

本稿の主定理を以下に与える．

定理 2. マッピングシステム $\mathcal{M}$ に関して，以下の条件は互いに同値である．

- (1) $\mathcal{M}$ は推移的である．
- (2) $\mathcal{E}$ を $\mathcal{M}$ に関する任意の評価システムとする時， $\mathcal{M}$ と $\mathcal{E}$ から導出されるマッピングカーネルは半正定値である．

(1) から (2) が導かれることは，Haussler による定理 1 の系として証明することができる．しかしながら，本稿では，逆に，定理 1 を定理 2 の特殊なケースと考えたいので，定理 1 を仮定せず，独立に定理 2 の証明を与える．実際， $M_{x, y} = \chi'_x \times \chi'_y$ は自明に推移性を満足する．

3. 類似性指標の分布に基づく類似性関数

この節では，マッピングカーネルの応用として，既知の類似性の指標から導かれる二つのカーネルを見る．従来の手法では指標の示す値のピーク(最大値及び最小値)

を類似度関数とするのに対し, この章で提案するカーネルは指標の値の分布を評価する点で新規である.

3.1 指標構造のサイズの分布に基づくカーネル

構造をもつ二つのデータオブジェクト (x, y) からある構造 t が共通に導出される時, 導出された構造はデータオブジェクトの類似性に関する情報を担っていると考えられる. このような (共通) 構造を, 指標構造 (*index structures*) と呼ぶ. x と y の類似度が高くなれば, より大きな指標構造がより多く存在すると考えられるので, 指標構造のサイズの分布を評価する類似度関数を式(5)で定義する. $IS(x, y)$ は x と y の指標構造 t の集合, $f: \mathbb{N} \rightarrow \mathbb{R}_+ = \{y \geq 0 \mid y \in \mathbb{R}\}$ は任意の正値関数とする.

$$K(x, y) = \sum_{t \in IS(x, y)} f(\text{size_of}(t)) \quad (5)$$

以下では, 木データの場合を例として, *agreement subtree* と *minimum refinement* の二つの指標構造を考える. この節では, 根付き (rooted) 順序付き (ordered) の木のみを扱う. 即ち, 木の頂点 (ノード) 間には, 根から葉へ階層による半順序と, 左右の位置による半順序が与えられているものとする.

木 t の木 x への埋入とは, t の頂点から x の頂点への単射で, *nearest common ancestor*^{*1} と左右の順序を保存しているものをさし, また, t が x にも y にも埋入されている時, t は x と y との間の *agreement subtree* (AST) であるという (図 1, [Berry 04]). 一般に AST は複数存在するが, その最大サイズを x と y の類似度を測る尺度として利用出来ることが知られている. MAST (*maximum agreement subtrees*) 問題とは, 与えられた木のペアに対して, 最大のサイズをもつ AST を少なくともひとつ探索する問題と定式化される. MAST 問題は広い応用範囲をもつことから (evolutionary trees [Hein 96, Berry 04], shape-axis trees [Pelillo 02] 等), 特にアルゴリズムの観点から精力的に研究されてきた.

MAST 問題では最大 AST のみを取り扱い, より小さいサイズの AST を全く無視する. 対照的に, 分布情報を評価することで, 式(5)のカーネルにはより高い分類性能を期待することができる. 以下では, $AST(x, y)$ は x と y の間の AST の集合を表すとし, $IS(x, y) = AST(x, y)$ として類似度関数 $K_{AST}(x, y)$ を定義する. $\text{size_of}(t)$ は t の大きさを表す関数であり, 葉の数 [Berry 04], ラベル付き頂点の数 [Kao 07] などの定義がある.

一般に, $K_{AST}(x, y)$ が効率的に計算できるとは限らないが, $f(n) = \alpha^n$ または $f(n) = n$ とする時には, 計算量が $O(|x| \cdot |y|)$ の与えられる動的アルゴリズムが存在する ($|x|$ は木 x の頂点の数, [Shin 08]).

図 1 Agreement subtree

更に, $K_{AST}(x, y)$ を以下のようにマッピングカーネルと見なす時, $\{M_{x, y}\}$ は推移的であり, $k(x', y')$ も半正定値であるので, 定理 2 により $K(x, y)$ が半正定値であることが導かれる.

- χ'_x は x の部分木の全体とする.
- $M_{x, y}^{AST} = \{(x', y') \in \chi'_x \times \chi'_y \mid x' \cong y'\}$. 但し, $x' \cong y'$ は, 同一の AST の埋入の像であることを示す.
- $k(x', y') = \begin{cases} f(\text{size_of}(x')) & \text{if } \text{size_of}(x') = \text{size_of}(y'), \\ 0 & \text{otherwise.} \end{cases}$

$$K_{AST}(x, y) = \sum_{(x', y') \in M_{x, y}^{AST}} k(x', y')$$

AST に加えて, *subtree isomorphism* [Pelillo 02, Aoki 03] や *agreement supertree* [Jansson 05] 等も木データの指標構造として利用可能で, マッピングシステムは推移的となるので同様に半正定値カーネルを導く.

一方, *minimum refinement subtree* [Hein 96, Berry 04] も指標構造として利用可能であるが, 必ずしも半正定値カーネルを導かない. 木のペア x と y と, 部分木のペア $x' \subseteq x$ と $y' \subseteq y$ に対して, x' と y' とが共に共通の木 t から辺 (エッジ) の縮約 (*edge contractions*) によって得られるとき, t は x' と y' の *refinement* であるといい, 頂点の数が最小の *refinement* を *minimum refinement* という. x' と y' とに対して, *minimum refinement* は一意に定まる. $MRT(x, y)$ を x 及び y の部分木ペアに対する *minimum refinement* の集合とし, $IS(x, y) = MRT(x, y)$ として式(5)から得られる類似度関数を $K_{MRT}(x, y)$ とする. 更に, マッピングシステム $M_{x, y}^{MRT}$ を以下のように定めると, $K_{MRT}(x, y)$ はマッピングカーネルとなる.

$$M_{x, y}^{MRT} = \{(x', y') \mid \exists (t \in MRT(x, y)) [x' \text{ と } y' \text{ は } t \text{ の像}]\}$$

しかしながら, $M_{x, y}^{MRT}$ は必ずしも推移的ではない. 実際, 図 2 において, s は x 及び y の, t は y 及び z の *minimum refinement* となるが, x と z の間には *minimum refinement* は存在しない. 即ち, $(x, y) \in M_{x, y}^{MRT}$ 及び $(y, z) \in M_{y, z}^{MRT}$

*1 最も近い共通の祖先

図2 Minimum refinement

が成り立つが, $(x, z) \notin M_{y,z}^{\text{MRT}}$ となってしまう. 更に, $f(n)$ を恒等的に 1 とすると, $K_{\text{MRT}}(x, x) = K_{\text{MRT}}(y, y) = K_{\text{MRT}}(z, z) = K_{\text{MRT}}(x, y) = K_{\text{MRT}}(y, z) = 7$ 及び $K_{\text{MRT}}(x, z) = 6$ より, 下記のように Gram 行列の行列式は負となり, $K_{\text{MRT}}(x, y)$ は半正定値ではない.

$$\det \begin{bmatrix} 7 & 7 & 6 \\ 7 & 7 & 7 \\ 6 & 7 & 7 \end{bmatrix} = -7 < 0$$

3.2 編集距離の分布に基づくカーネル

編集距離は離散構造の間の類似度を測る尺度であり, 文字列 ([Wagner 74] 等), 木 ([Barnard 95] 等), 一般のグラフ ([Bunke 97] 等) 等, 広い範囲で用いられている.

まず, 編集距離の汎用的な定義を述べる. x を一つ以上の要素を含む構造であるとする. 例えば, 文字列は文字を要素とする構造であり, グラフは頂点と辺を要素とする構造である. 編集操作(edit operations)は, x の要素に対する操作であり, (i) 要素 a の要素 b への置き換え ($\langle a \rightarrow b \rangle$), (ii) 要素 a の削除 ($\langle a \rightarrow \bullet \rangle$), (iii) 要素 a の挿入 ($\langle \bullet \rightarrow a \rangle$) のいずれかである. 編集操作列(edit scripts)とは, 0 個以上の編集操作の列であり, 構造を他の構造への変換するスクリプトであると解される. また, 個々の編集操作 $\langle a \rightarrow b \rangle$ に対して編集コスト $\gamma\langle a \rightarrow b \rangle \in \mathbb{R}$ を定義し, 編集操作列 σ のコスト $\gamma(\sigma)$ を σ を構成する編集操作のコストの和として定義する. 以上の定義のもとで, 構造 x と y の間の編集距離 $d(x, y)$ は,

$$d(x, y) = \min\{\gamma(\sigma) \mid \sigma \text{ は } x \text{ を } y \text{ に変換する編集操作列}\}$$

と定義される. 確かに, 編集距離は構造間の類似度を測る尺度として有効であるが, 3.1 節の場合と同じように, 必ずしも最小コストを与えない編集操作列の分布を無視している嫌いがある. そこで, この節では, 式(6)に従っ

て, 編集操作列のコストの分布を評価するカーネルを導入する. $f: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ は任意の正值関数である.

$$K(x, y) = \sum_{\sigma \text{ は } x \text{ を } y \text{ に変換}} f(\gamma(\sigma)) \quad (6)$$

本稿では, 更に以下を仮定する.

- (1) コストは対称である ($\gamma\langle \xi \rightarrow \eta \rangle = \gamma\langle \eta \rightarrow \xi \rangle$).
- (2) $f(x) = e^{-x}$ とする*2.
- (3) $\sigma = \langle x'_1 \rightarrow y'_1 \rangle \dots \langle x'_n \rightarrow y'_n \rangle$ は既約である, 即ち, 以下の条件を満足する*3.
 - a $x'_1, \dots, x'_n$ には, x の各要素が丁度 1 回ずつ現れ, その他は $\bullet$ である.
 - b $y'_1, \dots, y'_n$ には, y の各要素が丁度 1 回ずつ現れ, その他は $\bullet$ である.
- (4) σ から挿入と削除を除外して得られる編集操作列を $\langle x'_{i_1} \rightarrow y'_{i_1} \rangle \dots \langle x'_{i_m} \rightarrow y'_{i_m} \rangle$ とする時, σ のトレースを $\text{tr}(\sigma) = ((x_{i_1}, \dots, x_{i_m}), (y_{i_1}, \dots, y_{i_m}))$ で定める. トレースが一致する編集操作列は, 構成する編集操作の順序を除いて一意である. これらの編集操作列を同一視し, 特に, 式(6)の中で一回だけ評価する. 上記の条件の下では, 式(6)のカーネル $K(x, y)$ の半正定値性を, 以下のように導くマッピングカーネル $\bar{K}(x, y)$ の半正定値性に帰着することが出来る.

$$g(x) = \prod_{\xi \in x} e^{-\gamma\langle \xi \rightarrow \bullet \rangle}, \quad g(y) = \prod_{\eta \in y} e^{-\gamma\langle \eta \rightarrow \bullet \rangle}$$

$$M_{x,y} = \{\text{tr}(\sigma) \mid \sigma: x \text{ から } y \text{ への編集操作列}\} \quad (7)$$

$$k(x', y') = \prod_{j=1}^m \frac{e^{-\gamma\langle \xi_j \rightarrow \eta_j \rangle}}{e^{-\gamma\langle \xi_j \rightarrow \bullet \rangle} e^{-\gamma\langle \bullet \rightarrow \eta_j \rangle}} \quad (8)$$

$$\text{但し, } (x', y') = ((\xi_1, \dots, \xi_m), (\eta_1, \dots, \eta_m))$$

$$K(x, y) = g(x) \cdot g(y) \cdot \left(\sum_{(x', y') \in M_{x,y}} k(x', y') \right) \quad (9)$$

式(9)の右辺の最後の因子を $\bar{K}(x, y)$ と表すと, $\bar{K}(x, y)$ はマッピングカーネルであり, 更に, $K(x, y)$ が半正定値であることと $\bar{K}(x, y)$ が半正定値であることは同値である ($g(x)$ は 0 を値にとらない)*4.

式(8)は必ずしも半正定値カーネルを与えるとは限らないが, 一般的な編集距離の設定では半正値となる. 式(8)を以下のように展開すると, $e^{-\gamma\langle \xi \rightarrow \eta \rangle}$ が要素の集合の上

*2 [Li 05] や [Salzberg 97] における生成モデルに従って, コスト $\gamma\langle a \rightarrow b \rangle$ を要素 a が要素 b に置き換わる確率の対数の負値と定義する時,

$$f(\gamma(\sigma)) = e^{-\sum_{i=1}^n \log \Pr(x'_i \rightarrow y'_i)} = \prod_{i=1}^n \Pr(x'_i \rightarrow y'_i)$$

が成立するので, $K(x, y)$ は x が y に変換される確率となる.

*3 任意の編集操作列は, 変換結果を変えることなく, 既約なものに縮退可能である.

*4 g によらず $g(x)g(y)$ は半正定値である.

で半正定値カーネルを与えることを示せば、式(8)は半正定値となる。

$$\begin{aligned} & \prod_{j=1}^m \frac{e^{-\gamma\langle \xi_j \rightarrow \eta_j \rangle}}{e^{-\gamma\langle \xi_j \rightarrow \bullet \rangle} e^{-\gamma\langle \bullet \rightarrow \eta_j \rangle}} \\ &= \left(\prod_{j=1}^m e^{\gamma\langle \xi_j \rightarrow \bullet \rangle} \right) \cdot \left(\prod_{j=1}^m e^{\gamma\langle \bullet \rightarrow \eta_j \rangle} \right) \cdot \left(\prod_{j=1}^m e^{-\gamma\langle \xi_j \rightarrow \eta_j \rangle} \right) \\ &= (h(x') \cdot h(y')) \cdot \left(\prod_{j=1}^m e^{-\gamma\langle \xi_j \rightarrow \eta_j \rangle} \right) \end{aligned}$$

Hamming 距離や Levenshtein 距離等、主な編集距離では、 ξ と η のラベルが等しい時には $\gamma\langle \xi \rightarrow \eta \rangle = 0$ 、ラベルが異なる場合には $\gamma\langle \xi \rightarrow \eta \rangle = 1$ と定めるが、この場合には $e^{-\gamma\langle \xi \rightarrow \eta \rangle}$ は半正定値となる*5。

式(8) が半正定値カーネルを与える時、定理 2 より、式(7) で定められるマッピングシステムが推移的であれば、 $\bar{K}(x, y)$ は半正定値となる。以下では、文字列と木データを例にとって、 $M_{x, y}$ の推移性を調べる。

文字列では Hamming 距離と Levenshtein 距離が代表的な編集距離である。 X_i 及び Y_i を文字とし、二つの文字列 $x = X_1 X_2 \dots X_k$ と $y = Y_1 Y_2 \dots Y_l$ を考える。マッピングシステム $M_{x, y}$ は、Hamming 距離では式(10)で、Levenshtein 距離では式(11)で与えられる。

$$M_{x, y} = \bigcup_{a=0}^{\min\{k, l\}} \{((X_{i_1}, \dots, X_{i_a}), (Y_{j_1}, \dots, Y_{j_a})) \mid 0 < i_1 < \dots < i_a \leq \min\{k, l\}\} \quad (10)$$

$$M_{x, y} = \bigcup_{a=0}^{\min\{k, l\}} \{((X_{i_1}, \dots, X_{i_a}), (Y_{j_1}, \dots, Y_{j_a})) \mid 0 < i_1 < \dots < i_a \leq k \wedge 0 < j_1 < \dots < j_a \leq l\} \quad (11)$$

いずれのマッピングシステムも推移的である。

木に関しては、最も広義の Tai 編集距離と less constrained 編集距離の二つを考察する。 x と y を根付き木 (rooted tree) とし、 $x' = (x'_1, \dots, x'_m)$ と $y' = (y'_1, \dots, y'_m)$ はそれぞれ x と y の頂点の一部からなる列を表すものとする時、Tai 編集距離、less constrained 編集距離の $M_{x, y}$

*5 簡単のために、 $\xi_1, \dots, \xi_n$ は互いに異なるラベルを有するとする時、 $e^{-\gamma\langle \xi_i \rightarrow \xi_j \rangle}$ に関する Gram 行列 G は以下のように二つの半正定値行列の和と表されるので、 G も半正定値となる。

$$G = \begin{bmatrix} 1 - e^{-1} & & \mathbb{O} \\ & \ddots & \\ \mathbb{O} & & 1 - e^{-1} \end{bmatrix} + \begin{bmatrix} e^{-1} & \dots & e^{-1} \\ \vdots & \ddots & \vdots \\ e^{-1} & \dots & e^{-1} \end{bmatrix}$$

$$\begin{aligned} ((A, B, C), (A, B, C)) &\in M_{x, y} \\ ((A, B, C), (A, B, C)) &\in M_{y, z} \\ ((A, B, C), (A, B, C)) &\notin M_{x, z} \end{aligned}$$

図3 Less-constrained 編集距離

は、各々、式(12)及び式(13)で与えられる。

$$M_{x, y} = \{(x', y') \mid x'_i < x'_j \Leftrightarrow y'_i < y'_j\} \quad (12)$$

$$\begin{aligned} M_{x, y} = \{(x', y') \mid [x'_i < x'_j \Leftrightarrow y'_i < y'_j] \wedge \\ [x'_i \sim x'_j < x'_i \sim x'_k \Rightarrow y'_i \sim y'_k = y'_j \sim y'_k]\} \end{aligned} \quad (13)$$

但し、 $x'_i < x'_j$ は x'_j が x'_i の祖先であることを表し、 $x'_i \sim x'_j$ は x'_i と x'_j の nearest common ancestor を表す。Tai 編集距離の $M_{x, y}$ は推移的であるのに対し、less constrained 編集距離の $M_{x, y}$ は推移的でない(図3)。

4. 定理 2 の証明

4.1 鍵となる補題

X^{ij} を添字 $(i, j) = \{1, \dots, n\}^2$ でパラメータ付けられた m 次元正方行列であるとし、 $X = [X^{ij}]_{i, j=1, \dots, n}$ を mn 次元正方行列で、 $(m(i-1) + g, m(j-1) + h)$ 成分が X^{ij} の (g, h) 成分 X_{gh}^{ij} と一致するものとする。

更に、 m 次元正方行列 A に対し、 $\text{smry}_A(X)$ を n 次元正方行列で、 $[\text{tr}(A^T X^{ij})]_{i, j=1, \dots, n}$ とする。 $\text{smry}_A(X)$ の (i, j) 成分は式(14)で与えられる。

$$\text{tr}(A^T X^{ij}) = \sum_{g=1}^m \sum_{h=1}^m A_{gh} X_{gh}^{ij} \quad (14)$$

補題 1. m 次元正方行列 A について、以下は同値。

1. A は半正定値
2. 任意の mn 次元半正定値行列 X に対して、 $\text{smry}_A(X)$ は半正定値

証明 まず、 A が対角行列である場合に、証明を行う。 A の (I, I) 成分を α_I と表す。

2 から 1 は、 $n = 1$ とし、更に、 X を、 $X_{II} = 1$ 、かつ、他の成分は 0 となるように定めることにより $\alpha_I \geq 0$ が得られることによる。

逆に、1 から 2 は、式(15)に従う。即ち、 $X = Y^T Y$ となる mn 次元正方行列 Y に対し、 $(m^2 n)$ 行 n 列行列 Z を $Z_{mn(I-1) + m(g-1) + i, j} = \sqrt{\alpha_I} Y_{iI}^{gj}$ となるように定める

と, $\text{smry}_A(X) = Z^T Z$ が成り立つ.

$$\begin{aligned} \text{tr} A^T X^{ij} &= \sum_{I=1}^m \alpha_I \left(\sum_{g=1}^n \sum_{h=1}^m Y_{hI}^{gi} Y_{hI}^{gj} \right) \quad (15) \\ &= \sum_{I=1}^m \sum_{g=1}^n \sum_{h=1}^m (\sqrt{\alpha_I} Y_{hI}^{gi}) (\sqrt{\alpha_I} Y_{hI}^{gj}) \end{aligned}$$

A が対角でない一般的なケースは, 以下のように, 対角行列のケースに帰着される. 即ち, P を $P^T A P$ が対角となるようにとってくる時, $\tilde{X} = [P^T X^{ij} P]_{i,j=1,\dots,n}$ に対して, $\text{smry}_A(X) = \text{smry}_{P^T A P}(\tilde{X})$ が成立する. ■

4.2 証明の準備

K の半正定値性は χ に含まれる任意の有限集合に対する Gram 行列の半正定値性として定義されるので, χ は最初から有限集合 $\{x_1, \dots, x_n\}$ であり, χ'_{x_i} も有限集合としてよい. また, 評価システム $(\chi', k, \{\gamma_x\})$ の χ' に特殊な要素 $\bullet$ を追加して, 任意の $x' \in \chi'$ に対して $k(\bullet, \bullet) = k(\bullet, x') = k(x', \bullet) = 0$ と定義しても, k の半正定値性は損なわれない.

$\bar{\chi}'$ を直和集合 $\bigsqcup_{i=1}^n \chi'_{x_i}$, $\bar{x}'$ を $\bar{\chi}'$ 中で $x' \in \chi'_{x_i}$ に対応する要素, $\bar{M} = \{(\bar{x}', \bar{y}') \mid x, y \in \chi \wedge (x', y') \in M_{x,y}\}$ とした上で, 写像 $\bar{\gamma}_x: \bar{\chi}' \rightarrow \chi'$ を次のように定義する.

$$\bar{\gamma}_x(\bar{x}') = \begin{cases} \gamma_x(x') & \text{if } x' \in \chi'_x, \\ \bullet & \text{otherwise} \end{cases}$$

この時, $\mathcal{M}$ と $\mathcal{E}$ に関するマッピングカーネル K は, 以下のように書き直すことができる.

$$K(x, y) = \sum_{(\bar{x}', \bar{y}') \in \bar{M}} k(\bar{\gamma}_x(\bar{x}'), \bar{\gamma}_y(\bar{y}'))$$

$\bar{\chi}' = \{\bar{x}'_1, \dots, \bar{x}'_m\}$ として, m 次元行列 A と X^{ij} を

$$A_{gh} = \begin{cases} 1 & \text{if } (\bar{x}'_g, \bar{x}'_h) \in \bar{M}, \\ 0 & \text{otherwise.} \end{cases}$$

$$X_{gh}^{ij} = k(\bar{\gamma}_{x_i}(\bar{x}'_g), \bar{\gamma}_{x_j}(\bar{x}'_h))$$

と定めると, $K(x_i, x_j) = \text{tr}(A^T X^{ij})$ が成り立つ.

4.3 定理 2 の証明-(1) から (2)

補題 1 により, A が半正定値であることを示せば十分である. ($X = [X^{ij}]_{i,j=1,\dots,n}$ は定義より半正定値である). A が対称であることは, マッピングシステムの定義より明らかである. また, $\{M_{x,y}\}$ が推移的であることから, $\{1, \dots, m\} = U_1 \sqcup \dots \sqcup U_\mu$ と直和分解し, $(\bar{x}'_g, \bar{x}'_h) \in \bar{M}$ が成り立つことと $g, h \in U_a$ となる $a \in \{1, \dots, \mu\}$ が存在することとが同値であるようにすることができる. 従って, $A = \bigoplus_{a=1}^\mu A[U_a]$ が成り立つ. $A[U_a]$ は半正定値行列なので A も半正定値である.

4.4 定理 2 の証明-(2) から (1)

対偶を証明する. $\mathcal{M}$ が推移的でないとすると, A は式(16) から(18) のうちの少なくとも一つを部分行列としてもつ. 下式では, 一般性を失うことなく, $g < h < b$ を仮定する.

$$A[g, h] = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad (16)$$

$$A[g, h] = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \quad (17)$$

$$A[g, h, b] = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \quad (18)$$

これら部分行列の行列式は負であり, 従って, 半正定値とはならない.

証明の方針は, 式(16) 乃至(18) で与えられる部分行列が存在するとして, χ に関する Gram 行列である $\text{smry}_A(X)$ が半正定値とならないような評価システム $\mathcal{E} = (\chi', k, \{\gamma_x\})$ が存在することを示すことである. 本稿では式(18) の場合のみを示すが, 残りの場合も同様に示すことができる.

i, j 及び a を, $x'_g \in \chi'_{x_i}$, $x'_h \in \chi'_{x_j}$ 及び $x'_b \in \chi'_{x_a}$ が成り立つように選ぶ. $\bar{\chi}'$ は, χ'_x ($x \in \chi$) の直和集合なので, このような i, j 及び a は一意に存在する. また, i, j 及び a は, 必ずしも互いに異なるとは限らない.

次に, 互いに直交する列ベクトル $\vec{e}_1, \vec{e}_2$ 及び $\vec{e}_3$ を, 以下の性質が成立するように選ぶ.

$$[\vec{e}_1, \vec{e}_2, \vec{e}_3]^T A[g, h, b] [\vec{e}_1, \vec{e}_2, \vec{e}_3] = \begin{bmatrix} \alpha_1 & 0 & 0 \\ 0 & \alpha_2 & 0 \\ 0 & 0 & \alpha_3 \end{bmatrix}$$

行列 $A[g, h, b]$ は半正定値ではないので, 一般性を失うことなく $\alpha_1 < 0$ としてもよい. 更に,

$$G = [\vec{e}_1, \vec{e}_2, \vec{e}_3] \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} [\vec{e}_1, \vec{e}_2, \vec{e}_3]^T$$

で定められる行列 K は半正定値であり,

$$\begin{aligned} &\text{tr}(A[g, h, b]^T K) \\ &= \text{tr} \left(\begin{bmatrix} \alpha_1 & 0 & 0 \\ 0 & \alpha_2 & 0 \\ 0 & 0 & \alpha_3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \right) = \alpha_1 < 0 \end{aligned}$$

を満たす.

$\mathcal{E} = (\chi', k, \{\gamma_x\})$ を以下を満足するように定める.

- $\chi' = \{\bullet, \xi, \eta, \zeta\}$
- $\begin{bmatrix} k(\xi, \xi) & k(\xi, \eta) & k(\xi, \zeta) \\ k(\eta, \xi) & k(\eta, \eta) & k(\eta, \zeta) \\ k(\zeta, \xi) & k(\zeta, \eta) & k(\zeta, \zeta) \end{bmatrix} = K$

$$\bullet \gamma_x(x') = \begin{cases} \xi, & \text{if } x = x_i \text{ and } x' = x'_g, \\ \eta, & \text{if } x = x_j \text{ and } x' = x'_h, \\ \zeta, & \text{if } x = x_a \text{ and } x' = x'_b, \\ \bullet, & \text{otherwise.} \end{cases}$$

以下では, 3つのケースに分けて, それぞれ, $\text{smry}_A(X)[i]$, $\text{smry}_A(X)[i, a]$, $\text{smry}_A(X)[i, j, a]$ が半正定値でないことを示し, もって, $\text{smry}_A(X)$ が半正定値でないことを証明する (定義 1, (5)).

$i = j = a$ の時: $\text{smry}_A(X)[i] = \text{tr}(A[g, h, b]^T K) < 0$ より, $\text{smry}_A(X)[i]$ は半正定値ではない (定義 1, (5)).

$i = j \neq a$ の時: 式(19)より, $\text{smry}_A(X)[i, a]$ は半正定値ではない (補題 1).

$$\text{tr}(A^T X^{ii}) = \text{tr}(A[g, h, b]^T [1, 2] K [1, 2])$$

$$\text{tr}(A^T X^{ia}) = A[g, h, b]^T_{1,3} K_{1,3} + A[g, h, b]^T_{2,3} K_{2,3}$$

$$\text{tr}(A^T X^{ai}) = A[g, h, b]^T_{3,1} K_{3,1} + A[g, h, b]^T_{3,2} K_{3,2}$$

$$\text{tr}(A^T X^{aa}) = A[g, h, b]^T_{3,3} K_{3,3}$$

$$\text{tr} \left(\text{smry}_A(X)[i, a] \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \right) = \text{tr}(A[g, h, b]^T K) < 0 \quad (19)$$

$i \neq j \neq a \neq i$ の時: 式(20)より, $\text{smry}_A(X)[i, j, a]$ は半正定値とはならない (補題 1).

$$\text{smry}_A(X)[i, j, a]_{\alpha, \beta} = A[g, h, b]^T_{\alpha, \beta} K_{\alpha, \beta}, \quad \alpha, \beta = 1, 2, 3$$

$$\text{tr} \left(\text{smry}_A(X)[i, j, a] \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \right) = \text{tr}(A[g, h, b]^T K) < 0 \quad (20)$$

5. 実験例

マッピングカーネルの利点は, 効果的な半正定値カーネルを広い範囲から選択できる点にある. このことを, 3.1 節で述べたカーネル $K_{\text{AST}}(x, y)$ と鹿島等 [Kashima 02, 鹿島 06] による elastic tree kernel との比較を例にとって確かめる. $K_{\text{AST}}(x, y)$ において, $\text{size_of}(t) = |t|$ 及び $f(n) = \alpha^n$ とし, パラメータ α を 0.1 から 2.0 まで 0.1 刻みに変動させて, $K_{\text{AST}}(x, y)$ の分類性能を調べる.

$$K_{\text{AST}}(x, y) = \sum_{(x', y') \in M_{x, y}^{\text{AST}}} \alpha^{\text{size_of}(x')}$$

Elastic tree kernel は, $\alpha = 1.0$ とした時の $K_{\text{AST}}(x, y)$ と一致する. 実験では, KEGG/GLYCAN データベース ([Hashimoto 06]) から取得した糖鎖のデータセットを使用し, LIBSVM を学習エンジンとした. 図 4 で示すように, $\alpha = 0.2$ の時に最もよい性能を示し, α が 0.2 より大きくなるに従って漸次性能が低下するという結果が得ら

図 4 糖鎖を用いた実験例

れている*6. このように, マッピングカーネルを利用することにより, 最適なパラメータ設定が可能となる.

◇ 参考文献 ◇

- [Aoki 03] Aoki, K. F., Yamaguchi, A., Okuno, Y., Akutsu, T., Ueda, N., Kanehisa, M., and Mamitsuka, H.: Efficient tree-matching methods for accurate carbohydrate database query, *Genome Informatics*, Vol. 14, pp. 134–143 (2003)
- [Barnard 95] Barnard, D., Clarke, G., and Duncan, N.: Tree-to-tree correction for document trees, *Technical Report 95-375*, Queen's University, Kingston, Ontario K7L 3N6 Canada (1995)
- [Berry 04] Berry, V. and Nicolas, F.: Maximum Agreement and Compatible Supertrees (Extended Abstract), *Proc. of the 5th Annual Symposium on Combinatorial Pattern Matching (CPM 2004)*, LNCS 3109, pp. 205–219 (2004)
- [Bunke 97] Bunke, H.: On a relation between graph edit distance and maximum common subgraph, *Pattern Recognition Letters*, Vol. 18, pp. 689–694 (1997)
- [Collins 01] Collins, M. and Duffy, N.: Convolution Kernels for Natural Language, *Proc. of Advances in Neural Information Processing Systems 14 (NIPS 2001)*, pp. 625–632, MIT Press (2001)
- [Cristianini 00] Cristianini, N. and Shawe-Taylor, J.: *An Introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge University Press (2000)

*6 χ で示される結果は, $f(n) = n$ の時の $K_{\text{AST}}(x, y)$ によるもので, elastic tree kernel とほぼ同等の性能を示した.

- [Hashimoto 06] Hashimoto, K., Goto, S., Kawano, S., Aoki-Kinoshita, K. F., and Ueda, N.: KEGG as a glycome informatics resource, *Glycobiology*, Vol. 16, pp. 63R – 70R (2006)
- [Haussler 99] Haussler, D.: Convolution Kernels on Discrete Structures, UCSC-CRL 99-10, Dept. of Computer Science, University of California at Santa Cruz (1999)
- [Hein 96] Hein, J., Jiang, T., Wang, L., and Zhang, K.: On the complexity of comparing evolutionary trees, *Discrete Applied Mathematics*, Vol. 71, pp. 153 – 169 (1996)
- [Jansson 05] Jansson, J., Ng, J. H. K., Sadakane, K., and Sung, W. K.: Rooted maximum agreement supertrees, *Algorithmica*, pp. 293 – 307 (2005)
- [Kao 07] Kao, M.-Y., Lam, T.-W., Sung, W.-K., and Ting, H.-F.: An Even Faster and More Unifying Algorithm for Comparing Trees via Unbalanced Bipartite Matchings, *Journal of Algorithms*, Vol. 40(2), pp. 212–233 (2007)
- [Kashima 02] Kashima, H. and Koyanagi, T.: Kernels for Semi-Structured Data., *Proc. of the 9th International Conference on Machine Learning (ICML 2002)*, pp. 291–298 (2002)
- [鹿島 06] 鹿島久嗣, 坂本比呂志, 小柳光生: 木構造データに対するカーネル関数の設計と解析, *人工知能学会論文誌*, Vol. 21, No. 1 (2006)
- [Kuboyama 06] Kuboyama, T., Shin, K., and Kashima, H.: Flexible tree kernels based on counting the number of tree mappings, *Proc. of 4th International Workshop on Mining and Learning with Graphs (MLG 2006)*, pp. 61–72, 2006.
- [Leslie 04] Leslie, C., Eskin, E., Cohen, A., Weston, J., and Noble, W. S.: Mismatch string kernels for discriminative protein classification, *Bioinformatics*, Vol. 20, No. 4 (2004)
- [Li 05] Li, H. and Jiang, T.: A class of edit kernels for SVMs to predict translation initiation sites in eukaryotic mRNAs, *Journal of Computational Biology*, Vol. 12, No. 6., pp. 702–718 (2005)
- [Lodhi 02] Lodhi, H., Shawe-Taylor, J., Cristianini, N., and Watkins, C. J. C. H.: Text classification using string kernels, *Journal of Machine Learning Research*, Vol. 2, pp. 419–444 (2002)
- [Lu 01] Lu, C. L., Su, Z.-Y., and Tang, G. Y.: A New Measure of Edit Distance between Labeled Trees, *Proc. of the 7th Annual International Conference on Computing and Combinatorics (COCOON 2001)*, LNCS 2108, pp. 338–348 (2001)
- [Menchetti 05] Menchetti, S., Costa, F., and Frasconi, P.: Weighted Decomposition Kernels, *Proc. of the 22nd International Conference on Machine Learning (ICML 2005)*, pp. 585–592 (2005)
- [Pelillo 02] Pelillo, M.: Matching free trees, maximal cliques, and monotone game dynamics, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 11, pp. 1535–1541 (2002)
- [Salzberg 97] Salzberg, S. L.: A method for identifying splice sites and translational start sites in eukaryotic mRNA, *Computer Applications in the Biosciences*, Vol. 13, No. 4, pp. 365–376 (1997)
- [Shin 08] Shin, K. and Kuboyama, T.: Kernels based on distributions of agreement subtrees, *Proc. of the 21st Australasian Joint Conference on Artificial Intelligence (AUS-AI 2008)*, LNCS 5360, pp. 236–246 (2008)
- [Suzuki 04] Suzuki, J., Isozaki, H., and Maeda, E.: Convolution Kernels with Feature Selection for Natural Language Processing Tasks, *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, pp. 119–126 (2004)
- [Wagner 74] Wagner, R. and Fischer, M.: The string-to-string correction problem, *JACM*, Vol. 21, No. 1, pp. 168–173 (1974)
- [Zien 00] Zien, A., Rätsch, G., Mika, S., Schölkopf, B., Lengauer, T., and Müller, K. R.: Engineering Support Vector Machine Kernels That Recognize Translation Initiation Sites, *Bioinformatics*, Vol. 16, No. 9, pp. 799 – 807 (2000)

[担当委員: 小野田 崇]

2008 年 10 月 23 日 受理

著 者 紹 介

申 吉浩

カーネギーメロン大学日本校教授・東京大学大学院理学系研究科数学専門課程博士課程単位取得退学 (1990)・富士ゼロックス株式会社主幹研究員, 東京大学先端科学技術研究センター特任助手を経て, 現職・博士 (工学) (東京大学)・暗号理論, 機械学習などの研究に従事。

久保山 哲二 (正会員)

学習院大学計算機センター准教授・東京大学生産技術研究所特任研究員・1992 年九州大学工学部情報工学科卒業・1997 年同大学院システム情報科学研究科情報システム学専攻単位取得退学・同年東京大学国際・産学共同研究センター助手・同助教を経て, 現職・博士 (工学) (東京大学)・パターン照合, 離散データ構造間の類似度設計, 機械学習などの研究に従事。