射にもとづく類似性理論

申吉浩

兵庫県立大学応用情報科学研究科
Graduate School of Applied Informatics, University of Hyogo

Abstract: We propose a generic framework to evaluate similarity of data.

1 はじめに

データの類似性は機械学習において最も基本的な概念である。例えば、クラスタリングは互いに類似したデータを集めることが目的であり、類別はデータの類似性に基づいて新しいデータのクラスを予測する。パターン抽出・認証は、複数のデータに共通に現れる類似領域を特定する。いずれにおいても、データの類似性の定量的評価が、機械学習における鍵となる。実際、文献中では、データの類似性を測る多くの手法が提案されてきた。例えば、カーネルは SVM を含む多変量解析の手法に用いられるが、本質は類似性関数である。編集距離は、構造化データの反類似性を測る反類似性関数である。

本論文のテーマは、文字列、木、グラフといった、構造をもつデータについて、その類似性を評価する手法を統一的に考えることにある。本論文はこのテーマに関する研究の端緒であって、代表的な類似性評価の手法である、編集距離、パターン抽出、カーネルの関係を部分的にでも明らかにすることを目的とする。例えば、編集距離問題とパターン抽出問題の間の双対性を示す。即ち、編集距離を求めることは、実は、二つのデータの間で最も類似性の高い共通パターンを見つけることと同じであることを示す。また、編集距離を計算したり、共通パターンを特定することは、類似性の分布においてそのピークを探索することであり、対照的に、カーネルを計算することは分布そのものの評価であることを見る。

類似性の評価指標を統一的な基盤の受けで考えることは、新しい手法を体系的に開発することにもつながる。例えば、文献中では、編集距離とパターン抽出は独立に研究されてきた。しかしながら、前述の双対性を理解できれば、それぞれの手法を交換することができる。例えば、本論文では、編集距離に関連して多重アラインメントを近似計算するアルゴリズムであるセンタースター法を、複数のデータの共通（類似）パターンを近似的に求める目的で利用できることを示す。

本論文中では、$\delta_{x,y}$ は Kronecker のデルタ関数を表すものとし、即ち、$x=y$ であれば $\delta_{x,y}=1$、それ以外では $\delta_{x,y}=0$ が成り立つ。

2 射

2.1 オブジェクトと要素

本論文では、一つかそれ以上のコンポーネントによって構成されるデータについて考える。簡便のため、そのようなデータを**オブジェクト**と呼び、オブジェクトを構成するコンポーネントを**要素**と呼ぶ。定式化のために、要素の空間を $\mathscr{Z}$、オブジェクトの空間を $\mathscr{W}$ とする。$X\in\mathscr{W}$ は有限個の要素から構成されるので、X を $\mathscr{Z}$ の有限部分集合とみなすことができる。

データのもつ個別の性質によって、オブジェクトの要素は特定の方法で構造化されている。例えば、一方向に並べられた文字は文字列を構成する。木は頂点の集合であるが、頂点の構造は少なくとも、グラフの観点、順序集合の観点、台数構造の観点の三つの異なる観点から理解することができる。即ち、$\mathscr{Z}$ 及び $\mathscr{W}$ は、オブジェクトの構造を正しく表現できるように定義されなければならない。以下では、例として、文字列を表現する $(\mathscr{Z}^S,\mathscr{W}^S)$、順序集合として木を表現する $(\mathscr{Z}^{T_p},\mathscr{W}^{T_p})$、半群として木を表現する $(\mathscr{Z}^{T_s},\mathscr{W}^{T_s})$、及び、一般的なグラフを表現する $(\mathscr{Z}^G,\mathscr{W}^G)$ を導入する。

例 1 (文字列)**.** $\mathscr{Z}^S=\Sigma\times\mathbb{N}$ とし、$\mathscr{W}^S$ の元 X は、$X\subset\mathscr{Z}^S$ で、$X=\{(\ell_1,1),(\ell_2,2),\ldots,(\ell_n,n)\}$ と表されるものとする。容易に、X をアルファベット Σ 上の文字列と考えることができる。

例 2 (順序集合としての木)**.** X を木の頂点集合を表すものとする。$v>w$ により、v は w の祖先であることを表す。この時、$(X,>)$ は半順序集合 (partially ordered set, poset) となる。更に、$(X,>)$ は、以下の条件を満足する。(p1) $v>u$ かつ $w>u$ ならば、$v>w$ か $v\le w$ かのいずれかが成り立つ。(p2) 頂点 $r\in X$ が存在し、$r\ge v$ が全ての $v\in X$ に対して成り立つ。反対に、もし半順序集合 $(X,>)$ が上記の条件を満足するならば、X は木とな

る。$(\mathscr{Z}^{T_p},>)$ を半順序数号とし、$\mathscr{W}^{T_p}$ は、$X \subseteq \mathscr{Z}^{T_p}$ であり、かつ、$(X,>)$ が条件 (p1) 及び (p2) を満足するものからなるとする。明らかに、$\mathscr{W}^{T_p}$ は木の集合となる。

例 3 (半群としての木)**.** 木 X を半群として見ることも可能である。$v \bullet w$ が、v と w の**直近共通祖先**を表す時、$(X,\bullet)$ は可換半群となる。即ち、$(u \bullet v) \bullet w = u \bullet (v \bullet w)$、及び、$u \bullet w = w \bullet v$ が任意の $u,v,w \in X$ に対して成り立つ。更に、$(X,\bullet)$ は、(s1) $u \bullet u = u$、及び、(s2) $|\{u \bullet v, v \bullet w, w \bullet u\}| \leq 2$ の二つの性質をもつ。反対に、可換半群 $(X,\bullet)$ が (s1) 及び (s2) を満足すれば、X は木となることを示すことができる。$(\mathscr{Z}^{T_s},\bullet)$ を可換半群、$\mathscr{W}^{T_s}$ は、部分半群 $(\mathscr{Z}^{T_s},\bullet)$ で、(s1) 及び (s2) を満足するものからなるとする。この時、$\mathscr{W}^{T_s}$ は木の集合である。

例 4 (グラフ)**.** $\mathscr{Z}^G$ を $V \cup E$、$E = V \times V$ とする。$\mathscr{W}^G$ は、$X \subset \mathscr{Z}^G$ で、かつ、$(v,w) \in X \cap E$ ならば、$v \in X \cap V$ かつ $w \in X \cap V$ となるものからなるとする。X はグラフであり、$X \cap V$ は頂点集合を、$X \cap E$ は辺集合を与える。

要素 $(\ell_i,i) \in \mathscr{Z}^S = \Sigma \times \mathbb{N}$ の ℓ_i はラベルと考えることができる。同様に、木及びグラフの頂点にはラベルが付与されることが多い。本論文では、$\ell : \mathscr{Z} \to \Sigma$ はラベルを与える写像であるとする。Σ はラベルのアフファベットである。

2.2 射

射 μ は有限集合 X から有限集合 Y への一対一部分写像である。即ち、部分集合 $X' \subseteq X$ をとると、$\mu : X' \to Y$ は一対一写像（単射）となる。この X' を $\mathrm{Dom}(\mu)$ で表し、$\mu(X') \subseteq Y$ を $\mathrm{Ran}(\mu)$ で表す。さらに、$|\mu|$ は $|\mathrm{Dom}(\mu)|$ を表すものとする。

X と Y の間の類似性を測るために、射の集合 $\mathscr{M}_{X,Y}$ を定める。$\mathscr{M}_{X,Y}$ を定める標準的な方法は、データの構造を保存する部分写像を集めることである。例えば、半群構造を例にとると、準同型射は $f(x \bullet y) = f(x) \bullet f(y)$ を満足する。

例 5 (Hamming 射)**.** $\mathscr{W}^S$ と $n \in \mathbb{N}$ において、$\mathscr{W}^S_n$ を $\mathscr{W}^S$ 中で $|X| = n$ を満足する文字列の集合とする。$X,Y \in \mathscr{W}^S_n$ について、Hamming 射 $\iota_{X,Y}$ は、$(\ell_i,i) \in X$ を $(\ell'_i,i) \in Y$ に対応づける。$\mathscr{M}^H_{X,Y} = \{\iota_{X,Y}\}$ とする。

例 6 (Levenshtein 射)**.** Levenshten 射 μ は順序を保存する。即ち、$X \in \mathscr{W}^S$、$Y \in \mathscr{W}^S$ とする時、$(\ell_i,i),(\ell_j,j) \in X$、$(\ell'_{i'},i') = \mu((\ell_i,i))$、$(\ell_{j'},j') = \mu((\ell_j,j)) \in Y$ 及び $i < j$ が成り立てば、$i' < j'$ が成り立つ。$\mathscr{M}^L_{X,Y}$ は X から Y への Levenshtein 射の全体を表すものとする。

例 7 (Taï射)**.** $X,Y \in \mathscr{W}^{T_p}$ に対して、Taï射 μ は頂点間の世代順序を保存する。即ち、$v,w \in \mathrm{Dom}(\mu)$ について、$v > w$ は $\mu(v) > \mu(w)$ と必要かつ十分である。$\mathscr{M}^T_{X,Y}$ は X から Y への Taï射の全体とする。

例 8 (合意部分木射)**.** 直近共通祖先演算子 $\bullet$ に関して、木 X の部分半群 X' **合意部分木**と呼ぶ。合意部分木 X' は $\bullet$ に関して閉じているので根をもつ。$X,Y \in \mathscr{W}^{T_s}$ を木とする時合意部分木射 μ は X の合意部分木 X' から Y への一対一準同型である。$\mathscr{M}^A_{X,Y}$ を合意部分木射の全体とする。

例 9 (Bunke 射)**.** グラフ $X,Y \in \mathscr{W}^G$ に対して、Bunke 射 μ は部分グラフ $X' \subseteq X$ から Y への一対一グラフ準同型である。$\mathscr{M}^B_{S,Y}$ を Bunke 射の全体とする。

$\mathrm{Dom}(\mu)$ と $\mathrm{Ran}(\mu)$ の形から $\mathscr{M}_{X,Y}$ を定める方法も有効である。次の例では、よく知られた MAST (Maximum Agreement SubTree) 問題と関連して、射の集合を定める。

例 10 (合同合意部分木射)**.** MAST 問題 [5] は二つ以上の木の最大共通合意部分木を求める問題である。合意部分木 $X' \subseteq X$ と $Y' \subseteq Y$ が共通であるとは、互いに同相であり、かつ、同相写像において対応する頂点が同じラベルをもつことをいう。したがって、$X,Y \in \mathscr{W}^{T_s}$ に対して、$\mathscr{M}^C_{X,Y} \subseteq \mathscr{M}^A_{X,Y}$ を以下のように定める：$\mu \in \mathscr{M}^A_{X,Y}$ であり、任意の $v \in \mathrm{Dom}(\mu)$ が $\mu(v)$ と同じラベルを持つ時、かつ、その時に限り、$\mu \in \mathscr{M}^C_{X,Y}$ とする。二つの木の間の MAST 問題は、$\arg\max\{|\mu| \mid \mu \in \mathscr{M}^C_{X,Y}\}$ に属する合同部分木射を見つけることに他ならない。

2.3 推移的射

射系とは $\{\mathscr{M}_{X,Y} \mid X,Y \in \mathscr{W}\}$ であって、(1) $id_X \in \mathscr{M}_{X,X}$; (2) $\mu \in \mathscr{M}_{X,Y}$ ならば、$\mu^{-1} \in \mathscr{M}_{Y,X}$. $id_X : X \to X$ は X の恒等写像である。

以下に定める**推移性**は射系の重要な性質であり、編集距離の三角不等式やカーネルの正定値性の条件である。

定義 1. 射系 $\{\mathscr{M}_{X,Y} \mid X,Y \in \mathscr{W}\}$ が**弱推移的**であるとは、すべての $X,Y,Z \in \mathscr{W}$、$\mu \in \mathscr{M}_{X,Y}$、$\nu \in \mathscr{M}_{Y,Z}$ に対して、$\mathrm{Ran}(\mu) = \mathrm{Dom}(\nu)$ が成り立つならば、$\nu \circ \mu \in \mathscr{M}_{X,Z}$ 成り立つことをいう。

定義 2. 射系 $\{\mathscr{M}_{X,Y} \mid X,Y \in \mathscr{W}\}$ が**推移的**であるとは、すべての $X,Y,Z \in \mathscr{W}$、$\mu \in \mathscr{M}_{X,Y}$、$\nu \in \mathscr{M}_{Y,Z}$ に対して、$\nu \circ \mu \in \mathscr{M}_{X,Z}$ が成り立つことをいう。

2.4 類似性指標

類似性指標 $\Phi_{X,Y} : \mathscr{M}_{X,Y} \to \mathbb{R}$ は、$(X,Y) \in \mathscr{W} \times \mathscr{W}$ に対して、射 $\mu \in \mathscr{M}_{X,Y}$ の $\mathrm{Dom}(\mu)$ と $\mathrm{Ran}(\mu)$ との間の類似性を定める。本論文で導入する方法論において、$\Phi_{X,Y}$

は X と Y との間の類似性評価の基礎となる。3 節で見るように、$\max\{\Phi_{X,Y}(\mu) \mid \mu \in \mathscr{M}_{X,Y}\}$ により類似性を評価することは、$\Phi_{X,Y}$ の利用法としては、最も直接的である。以下では、$\Phi_{X,Y}$ を単純に Φ と表記する。

Φ を定義するために、要素間の原始類似性指標 $\varphi : \mathscr{Z} \times \mathscr{Z} \to \mathbb{R}$ を利用する。特に、原始類似性指標から、Φ_+ と $\Phi_\times$ の二つの類似性指標が導かれる。

$$\Phi_+(\mu) = \sum_{(x,y)\in\mu} \varphi(x,y)$$
$$\Phi_\times(\mu) = \prod_{(x,y)\in\mu} \varphi(x,y)$$

φ が正定値であると仮定することで、多くの利点を得ることができる。正定値カーネルの最も重要な性質は、**再生核ヒルベルト空間** *reproducing kernel Hilbert spaces* (RKHS) が存在することである [1]: ヒルベルト空間 $\mathscr{H}$ への埋め込み写像 $\mathscr{Z} \ni x \mapsto \varphi_x \in \mathscr{H}$ が存在して、すべての $x,y \in \mathscr{Z}$ に対して、$\varphi(x,y) = \langle \varphi_x, \varphi_y \rangle_{\mathscr{H}}$ が成り立つ。$\langle \cdot,\cdot \rangle_{\mathscr{H}}$ は $\mathscr{H}$ における内積を表す。

3 最大類似性問題

二つのデータオブジェクトの類似性を決定する問題は、以下の最大類似性問題として、最適化問題として定式化される。

Maximum Similarity Measurement (MSM) 問題:
$X,Y \in \mathscr{W}$ に対して、$\max\{\Phi(\mu) \mid \mu \in \mathscr{M}_{X,Y}\}$ を求めよ。

4 編集距離と MSM 問題の双対性

まず、コスト関数 $\psi_+ : \mathscr{Z} \times \mathscr{Z} \to \mathbb{R}$ を原始的類似性指標 $\varphi : \mathscr{Z} \times \mathscr{Z} \to \mathbb{R}$ から定義する。このために、ギャップ要素 $\perp$ を $\mathscr{Z}$ に追加し、φ を $\varphi(x,\perp) = \varphi(\perp,y) = \varphi(\perp,\perp) = 0$ により拡張する。簡単のため、拡張後も、同じ記号 $\mathscr{Z}$、φ を用いる。コスト関数 ψ_+ は以下のように定義される。

$$\psi_+(x,y) = \frac{1}{2}\varphi(x,x) + \frac{1}{2}\varphi(y,y) - \varphi(x,y). \qquad (1)$$

$\varphi(x,y)$ をヒルベルト空間内の内積 $\langle x,y \rangle$ と見なした場合、$\psi_+(x,y) = \frac{1}{2}\|x-y\|^2$ は余弦定理より導かれる。従って、$\psi_+(\cdot,\cdot)$ はある距離の二乗の半分に相当する。

$\psi_+(x,\perp)$ と $\psi_+(\perp,y)$ は編集操作の削除と挿入に相当し、$\psi_+(x,y)$ は置換操作を意味する。Eq. (1) は次の式と同値である。

$$\varphi(x,y) = \psi_+(x,\perp) + \psi_+(y,\perp) - \psi_+(x,y). \qquad (2)$$

$\mathscr{M}$	定義	編集距離	データ構造
$\mathscr{M}^H_{X,Y}$	例 5	Hamming	文字列
$\mathscr{M}^L_{X,Y}$	例 6	Levenshtein	文字列
$\mathscr{M}^T_{X,Y}$	例 7	Taï	木
$\mathscr{M}^A_{X,Y}$	例 8	Accordant	木
$\mathscr{M}^B_{X,Y}$	例 9	Graph	グラフ

表 1: 文献中で知られている $(\mathscr{M},\Phi_+)$-距離

命題 1. *1.* 任意の $x,y \in \mathscr{Z}$ について、$\psi_+(x,y) = \psi_+(y,x)$ が成り立つ。

2. 任意の $x \in \mathscr{Z}$ について、$\psi_+(x,x) = 0$ が成り立つ。

3. 任意の $x \in \mathscr{Z}$ について、$\psi_+(x,\perp) = \psi_+(\perp,x) = \frac{1}{2}\varphi(x,x)$ が成り立つ。

4. ψ_+ が負定値であることと、φ が正定値であることとは同値である。

系 1. φ が正定値であるならば、$\mathscr{Z}$ は $\psi_+(\cdot,\cdot)^{\frac{1}{2}}$ に関して擬距離空間となる。.

例 11. 編集距離の計算では、同じラベルの置換操作のコストを 0 とし、その他のすべての編集操作のコストを 1 とすることが通常である。即ち, ψ_+ は $\psi_+(x,y) = 1 - \delta_{\ell(x),\ell(y)}$, $\psi_+(\perp,\perp) = 0$、及び、$\psi_+(x,\perp) = \psi_+(\perp,y) = 1$ により定義される。この ψ_+ は、Eq. (1) により、$\varphi(x,y) = \delta_{\ell(x),\ell(y)} + 1$ から導出される。勿論、ψ_+ は負定値であり、φ は正定値である。ただし、異なる要素が同じラベルを有することが許されるので、$\sqrt{\psi_+(\cdot,\cdot)}$ は必ずしも距離空間を与えない。

次いで、$X,Y \in \mathscr{W}$ に対して、$\mu \in \mathscr{M}_{X,Y}$ の編集コスト $\Psi_+(\mu)$ を以下のように定義する。

$$\Psi_+(\mu) = \sum_{x\in X\setminus \mathrm{Dom}(\mu)} \psi_+(x,\perp) + \sum_{y\in Y\setminus \mathrm{Ran}(\mu)} \psi_+(\perp,y) + \sum_{(x,y)\in\mu} \psi_+(x,y).$$

定義 3. $d_{\mathscr{M},\Phi_+}(X,Y) = \min\{\Psi_+(\mu) \mid \mu \in \mathscr{M}_{X,Y}\}$.

例 5、6、7、8、9 で述べた $\mathscr{M}$ と例 11 の φ_ℓ から導出される Φ_+ に対して、得られる $(\mathscr{M},\Phi_+)$-距離はすべて文献でよく知られている編集距離と一致する (表 1)。

定理 1 は編集距離問題と MSM 問題との関係を明らかにする。

定理 1 (双対性)**.** 以下の等式が成り立つ。

$$d_{\mathscr{M},\Phi_+}(X,Y) = \frac{1}{2}\left(\sum_{x\in X}\varphi(x,x) + \sum_{y\in Y}\varphi(y,y)\right) - \max_{\mu\in\mathscr{M}_{X,Y}} \Phi_+(\mu)$$

[11] は、文献中で双対性を最初に指摘した論文であり、木の編集距離 (TED) 問題とパターン抽出問題との間の双対性を指摘した。具体的には、accordant 距離を求める TED 問題の双対問題として、MAAST (Mostly Adjusted Agreement-Subtree) 問題を定義した。定理 1 は、上記の結果を著しく一般化したものである。

5 パターン抽出

射系 $\mathscr{M} = \{\mathscr{M}_{X,Y} \mid X,Y \in \mathscr{W}\}$ の別の見方として、パターン抽出問題において抽出したい構造のクラスを指定する手段と見ることができる。

MAST 問題は以下のように定式化できる。

MAST 問題:
$\{X_1,\ldots,X_n\} \subseteq \mathscr{W}^{T_s}$ とする。すべての $i \neq j \neq k \neq i$ に対して (1) $\mu_{ji} = \mu_{ij}^{-1}$ 及び (2) $\mathrm{Dom}(\mu_{ij}) = \mathrm{Dom}(\mu_{ik})$ が成り立つ条件のもとで、$|\mu_{12}|$ を最大にする $\{\mu_{ij} \in \mathscr{M}^C_{X_i,X_j} \mid i,j = 1,\ldots,n; i \neq j\}$ を求めよ。

ここで、φ_ℓ を原始類似性指標ととれば、$\Phi_+(\mu) = |\mu|$ が成り立ち、$|\mu|$ は類似性指標となる。これを一般化して、次のようにパターン抽出問題を定式化することができる。

$(\mathscr{M},\Phi)$-パターン抽出 (PE) 問題:
$\{X_1,\ldots,X_n\} \subseteq \mathscr{W}$ とする。すべての $i \neq j \neq k \neq i$ に対して (1) $\mu_{ij} = \mu_{ji}^{-1}$ 及び (2) $\mu_{ij} = \mu_{ik} \circ \mu_{kj}$ が成り立つという条件のもとで、$\sum_{i=1}^{n}\sum_{j=i+1}^{n} \Phi(\mu_{ij})$ を最大にする $\{\mu_{ij} \in \mathscr{M}_{X_i,X_j} \mid i,j = 1,\ldots,n; i \neq j\}$ を求めよ。

Proposition 2 により、φ_ℓ が原始類似性指標であるとき、MAST 問題は $(\mathscr{M}^C,\Phi+)$-パターン抽出問題である。

命題 2. $\{\mu_{ij} \in \mathscr{M}_{X_i,X_j} \mid i \neq j\}$ を $(\mathscr{M},\Phi)$-パターン抽出問題の最適解とする。この時、$\mathrm{Dom}(\mu_{ji}) = \mathrm{Dom}(\mu_{ki}) = \mathrm{Ran}(\mu_{ij}) = \mathrm{Ran}(\mu_{ik})$ がすべての $i \neq j \neq k \neq i$ に対してなりたつ。

定理 1 は、$(\mathscr{M},\Phi)$-パターン抽出問題を $n = 2$ の場合に解くことが、$(\mathscr{M},\Phi)$-距離問題をとくことと同値であることを示す。一方、多くの場合、$(\mathscr{M},\Phi)$-距離問題は編集距離問題として効率的に解けることが知られている。対照的に、$(\mathscr{M},\Phi)$-パターン抽出問題を $n \geq 3$ の場合に解くことは一般に NP 困難である。

この問題を避けるために、定理 1 は**センタースター法**の利用が可能であることを示唆している。センタースター法は文字列の最適多重アラインメントの近似を計算するためのアルゴリズムであり [3]、ペアワイズの編集距離が効率的に計算可能であれば、高速に多重アラインメントを計算する。近似アルゴリズムとしては、近似保証が定数 2 で抑えられるという、非常によい性質を持つ。

以下では、センタースター法を $(\mathscr{M},\Phi)$-パターン抽出問題に適用する方法を示す。多重アラインメントを求める際のセンタースター法の場合と同じく、**ピボット**の計算は効率的であることを仮定する。$\{X,X_1,\ldots,X_n\} \subseteq \mathscr{W}$ に対して、X のまわりの**ピボット**とは、$\mathrm{Dom}(\mu_1) = \cdots = \mathrm{Dom}(\mu_n)$ という条件のもとで、$S = \sum_{i=1}^{n} \Phi(\mu_i)$ を最大にする $(\bar{\mu}_1,\ldots,\bar{\mu}_n) \in \mathscr{M}_{X,X_1} \times \cdots \times \mathscr{M}_{X,X_n}$ のことである。S をピボットの**シグニチャ**と呼ぶ。

$(\mathscr{M},\Phi)$-パターン抽出問題の近似解を計算するための「センタースター法」は、以下のように、記述される。

センタースター法:
$X_1,\ldots,X_n \in \mathscr{W}$ が与えられているとする。

1. X_i のまわりのピボット $\bar{\mu}_{i1}\ldots,\bar{\mu}_{i,i-1},\bar{\mu}_{i,i+1},\ldots\bar{\mu}_{in}$ を計算し、S_i をそのシグネチャとする。
2. $k \in \arg\max\{S_i \mid i = 1,\ldots,n\}$ をとる。
3. $i \neq k$ 及び $j \neq k$ に対して、$\mu_{ki} = \bar{\mu}_{ki}$、$\mu_{ik} = \bar{\mu}_{ki}^{-1}$ 及び $\mu_{ij} = \bar{\mu}_{kj} \circ \bar{\mu}_{ki}^{-1}$ を計算する。

センタースター法で $(\mathscr{M},\Phi)$-パターン抽出問題を解く最大のメリットは、近似保証が定数で与えられるという点にある。

定義 4. 原始的類似性指標 $\varphi : \mathscr{Z} \times \mathscr{Z} \to \mathbb{R}$ が**正有界**であるとは、$\inf\{\varphi(x,y) \mid x,y \in \mathscr{Z}\} > 0$ かつ $\sup\{\varphi(x,y) \mid x,y \in \mathscr{Z}\} < \infty$ が成り立つことである。また、$\sup\varphi / \inf\varphi$ を $c(\varphi)$ で表す。

定理 2. $X_1,\ldots,X_n \in \mathscr{W}$ と $(\mathscr{M},\Phi)$-パターン抽出問題に対して、$\{\mu_{ij} \in \mathscr{M}_{X_i,X_j} \mid i,j = 1,\ldots,n, i \neq j\}$ はセンタースター方で求めた近似解であるとし、$\{\widehat{\mu}_{ij} \in \mathscr{M}_{X_i,X_j} \mid i,j = 1,\ldots,n, i \neq j\}$ を最適解であるとする。φ が正有界であり、かつ、$\mathscr{M}$ が推移的ならば、以下が成り立つ。

$$\sum_{i=1}^{n}\sum_{j=i+1}^{n} \Phi_+(\mu_{ij}) \geq \frac{1}{c(\varphi)} \sum_{i=1}^{n}\sum_{j=i+1}^{n} \Phi_+(\widehat{\mu}_{ij})$$

6 モーメントカーネル

実確率変数 X と X 上の確率分布 P が与えられた時、この確率分布の n 次モーメントは以下のように定義される。

$$m_n = \int_{-\infty}^{\infty} x^n P(x) dx.$$

特に、X の定義域が有限集合 $\{x_1,\ldots,x_n\}$ である時は、n 次モーメントは

$$m_n = \sum_{i=1}^{n} x_i^n p(x_i)$$

となる。モーメントは分布を記述する統計量である。実際、1 次モーメント m_1 は分布の平均に他ならず、$m_2 - m_1^2$ は分散を与える。

定義 5. X,Y を $\mathscr{W}$ に属するオブジェクトとする。射系 $\mathscr{M}_{X,Y}$ と類似性指標 Φ に対して、n **次モーメントカーネル**を $K_n(X,Y) = \sum_{\mu \in \mathscr{M}_{X,Y}} \Phi(\mu)^n$ によって定義する。

$K_0(X,Y) = |\mathscr{M}_{X,Y}|$ が成り立つ。$K_1(X,Y)/K_0(X,Y)$ は、$\mu \in \mathscr{M}_{X,Y}$ にわたる $\Phi(\mu)$ の分布の平均値であり、

$$K_2(X,Y)/K_0(X,Y) - (K_1(X,Y)/K_0(X,Y))^2$$

はその分散を与える。

モーメントカーネルを利用する重要な利点は、SVM を含む強力な変数関数理論による分析手法を使える点にある。そのために、定理 3 は重要な役割を果たす。

定理 3 ([9])**.** Φ を Φ_+ 又は $\Phi_\times$ とし、原始類似性指標 φ から導かれるとする。$\mathscr{M}$ が推移的で、かつ、φ が正定値ならば、$K_n(X,Y)$ も正定値である。

$K_n(X,Y)$ が正定値であるためには、$\mathscr{M}$ の推移性が根拠になっていることに注意されたい。

文献中でも、構造化データの分析にカーネルを用いる手法は精力的に研究 s れてきた。最初の重要な貢献は**畳み込みカーネル** [4] であり、集合 S と T とに対して、$K(S,T) = \sum_{(x,y)\in S\times T} k(x,y)$ と定義される。この時、$k(x,y)$ が正定値ならば、$K(S,T)$ も正定値である。申・久保山 [8] は畳み込みカーネルを一般化して、**マッピングカーネル**を導入した。マッピングカーネルにより、構造を持つデータに対する正定値カーネルの設計が著しく容易になった。これらの貢献に基づき、文献中で多くのカーネルが提案されている。

例えば、**全文字列カーネル** [7] は、文字列に対するカーネルとして非常によく知られているが、実は、$\mathscr{M}^L$ に対する 0 次モーメントカーネルであることがわかる。木に対しては、[2] が解析木カーネルを導入し、[6] が弾性カーネルを導入している。これらも 0 次モーメントカーネルの例であり、特に、弾性カーネルは $\mathscr{M}^A$ に対する 0 次モーメントカーネルである。文字列と木に対しては、他にも多くのカーネルが知られているが、筆者が知る限り、ほとんど全てがなんらかの射系 $\mathscr{M}$ に対する 0 次モーメントカーネルとなる。[10] では、高次モーメントカーネルを含む多様なカーネルの計算可能性について論じている。

また、定理 4 は、オブジェクトの類似性とモーメントカーネルの間の非常に興味深い関係を与える。

定理 4. $\Phi(\mu) > 0$ がすべての $\mu \in \mathscr{M}_{X,Y}$ に対して成り立てば、以下の関係が成り立つ。

$$\max\{\Phi(\mu) \mid \mu \in \mathscr{M}_{X,Y}\} = \lim_{n\to\infty} K_n(X,Y)^{1/n}.$$

7 結論

構造化データの間の類似性を統一的な方法で評価する共通のフレームワークを提案した。このフレームワークにより、編集距離、パターン抽出、カーネルの関連性を明確に説明することができる。特に、類似性の分布を評価する手法としてモーメントカーネルを提案し、かつ、文献中で知られている構造化データに対するカーネルのほとんどが 0 次モーメントカーネルの例となることを示した。

参考文献

[1] C. Berg, J. P. R. Christensen, and R. Ressel. *Harmonic Analysis on semigroups. Theory of positive definite and related functions*. Springer, 1984.

[2] M. Collins and N. Duffy. Convolution kernels for natural language. In *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001]*, pages 625–632. MIT Press, 2001.

[3] D. Gusfield. Efficient methods for multiple sequence alignment with guaranteed error bounds. *Bulletin of Mathematical Biology*, 55:141–154, 1993.

[4] D. Haussler. Convolution kernels on discrete structures. UCSC-CRL 99-10, Dept. of Computer Science, University of California at Santa Cruz, 1999.

[5] Ming-Yang Kao, Tak-Wah Lam, Wing-Kin Sung, and Hing-Fung Ting. An even faster and more unifying algorithm for comparing trees via unbalanced bipartite matchings. July 2007.

[6] H. Kashima and T. Koyanagi. Kernels for semi-structured data. In *the 9th International Conference on Machine Learning (ICML 2002)*, pages 291–298, 2002.

[7] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[8] K. Shin and T. Kuboyama. A generalization of Haussler's convolution kernel - mapping kernel. In *ICML 2008*, 2008.

[9] K. Shin and T. Kuboyama. Generalization of haussler's convolution kernel - mapping kernel and its application to tree kernels. *J. Comput. Sci. Technol*, 25(5):1040–1054, 2010.

[10] Kilho Shin. Partitionable kernels for mapping kernels. In *ICDM 2011*, pages 645–654, 2011.

[11] Kilho Shin. Tree edit distance and maximum agreement subtree. *Inf. Process. Lett.*, 115(1):69–73, 2015.