7構造化データに対する正定値カーネル

7.1ストリングカーネル

08t40721 全　太俊

7.1.1 ストリングデータ

- ストリング

  –アルファベットと呼ばれる有限集合$\sum$の元からなる有限列

  例　$\sum = \{a, b, c, ..., z\}$

  cat, head, xxyzpgaaなどはストリング

  –記号法

  $\sum^p$　:長さpのストリング全体

  $\sum^* = \bigcup_{p=0}^{\infty} \sum^p$:ストリング全体

7.1.1 ストリングデータ

−記号法(続き)

ストリング$s=s_1s_2...s_n$に対し

$|s|$はストリングの長さn

$s[i:j]$は第i文字から第j文字までの部分列

$s=s_1...s_n$, $t=t_1...t_m$に対し

s, t結合$st=s_1...s_nt_1...t_m$

7.1.2 p-スペクトラムカーネル

- p-スペクトラムカーネル
  - 長さpの部分列が2つのストリングに共通に出現する回数に基づくカーネル
  - カーネルk(s, t)

$$k_p(s,t) = \sum_{u \in \sum^p} \phi_u^p(s)\phi_u^p(t)$$

$\phi_u^p(s)$:uがsの部分列として出現する回数

$\phi_u^p(t)$:uがtの部分列として出現する回数

7.1.2 p-スペクトラムカーネル

- 3-スペクトラムカーネル

  s=statistics　t=pastapistan

表7.2　3-スペクトラムカーネルの特徴写像の例

	sta	tat	ati	tis	ist	sti	tic	ics	pas	ast	tap	api	pis	tan
$\phi(s)$	1	1	1	1	1	1	1	1	0	0	0	0	0	0
$\phi(t)$	2	0	0	0	1	0	0	0	1	1	1	1	1	1

⇒ $k_3(s,t)=1*2+1*1=3$

7.1.2 p-スペクトラムカーネル

- p-スペクトラムの計算
  - sの先頭からp部分列を順に取り出し、それがtから同様に取り出したp部分に等しいか順に調べる

$$h_p(x,y) = \begin{cases} 1 & x[1:p]=y[1:p]のとき \\ 0 & x[1:p]\neq y[1:p]のとき \end{cases}$$

$$k_p(s,t) = \sum_{i=1}^{|s|-p+1} \sum_{j=1}^{|t|-p+1} h_p(s[i:|s|], t[j:|t|])$$

計算量=0(p|s||t|)

- 接尾辞木(suffix tree)

計算量=0(p(|s|+|t|))

7.1.3 全部分列カーネル

- 全部分列カーネル
  - ギャップを許した部分列の出現回数を特徴ベクトル
  - 長さnの列$i=[i_1, i_2, \ldots, i_n]$に対し、長さnのストリング$s[i]$は以下のように定義する

    $s[i]=s_{i1}s_{i2}\ldots s_{in}$

  - カーネル$k_{all}(s, t)$

$$k_{all}(s,t) = \sum_{u \in \sum^*} \phi_u^{all}(s)\phi_u^{all}(t)$$

7.1.3 全部分列カーネル

- 全部分列カーネルの例

例1　s=ATGACTAC　　t=CATGCGATT　　u=ATGCA　→　ATGACTAC　CATGCGATT

例2　s=ATG　t=AGC

表7.3 s=ATG,t=AGCに対する全部分列特徴写像

	ε	A	T	G	C	AT	AG	AC	TG	GC	ATG	AGC
$\phi(t)$	1	1	1	1	0	1	1	0	1	0	1	0
$\phi(t)$	1	1	0	1	1	0	1	1	0	1	0	1

⇒ $k_{all}(s, t)=4$

7.1.3 全部分列カーネル

- 全部分列カーネルの計算
  -ストリングの共通部分列を直接すべて例挙するための計算量は非常に大きい
  -再帰的なアルゴリズム(1)-sのみに対する再帰

    k(s, ε)=k(ε, t)=1 ←――― 初期条件

    k(s. t)がすでに求められていると仮定

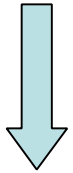
$$k(sa,t) = k(s,t) + \sum_{\substack{1 \le j \le |t| \\ j:t[j]=a}} k(s, t[1, j-1])$$

    計算量=0|s||t|^2

7.1.3 全部分列カーネル

- 再帰的アルゴリズム(2)−s, tに対する再帰

$$\tilde{k}(sa, t) = \sum_{\substack{1 \le j \le |t| \\ j : t[j] = a}} k(s, t[1, j-1])$$ とおく



1≤j≤|t|の場合とj=|tb|の場合に分けると

$$\tilde{k}(sa, tb) = \sum_{\substack{1 \le j \le |t| \\ j : t[j] = a}} k(sa, t) + \delta_{ab} k(s, t) = \tilde{k}(sa, t) + \delta_{ab} k(s, t)$$

$$\begin{cases} k(sa, t) = k(s, a) + \tilde{k}(sa, t) & \text{sに関する再帰式} \\ \tilde{k}(sa, tb) = \tilde{k}(sa, t) + \delta_{ab} k(s, t) & \text{tに関する再帰式} \end{cases}$$

計算量=0|s||t|

7.1.4 ギャップ重みつき部分列カーネル

- ギャップ重みつき部分列カーネル

  –ギャップの個数によって重みをつけた部分列による特徴ベクトル

  –カーネル

$$k_w(s,t) = \sum_{u \in \sum^p} \phi_u^{p,\lambda}(s)\phi_u^{p,\lambda}(t)$$

$$\phi_u^{p,\lambda}(s) = \sum_{i:s[i]=u} \lambda^{\|s[i]\|}$$

$i=(i_1,\ldots,i_n)$のとき$||s[i]||=i_n-i_1+1$

7.1.4 ギャップ重みつき部分列カーネル

- ギャップ重みつき部分列カーネルの例
  s=ATGC, t=AGCT, p=2

表7.4ギャップ重みつき部分列カーネルの特徴ベクトルの例

	AT	AG	AC	TG	TC	GC	GT	CT
$\phi(s)$	λ^2	λ^3	λ^4	λ^2	λ^3	λ^2	0	0
$\phi(t)$	λ^4	λ^2	λ^3	λ^4	0	λ^2	λ^3	λ^2

$$\Longrightarrow \quad k_w(s,t) = \lambda^4 + \lambda^5 + 2\lambda^6 + \lambda^7$$