

カーネル法による 構造データの解析

鹿島 久嗣
IBM 東京基礎研究所

© 2005 IBM Corporation

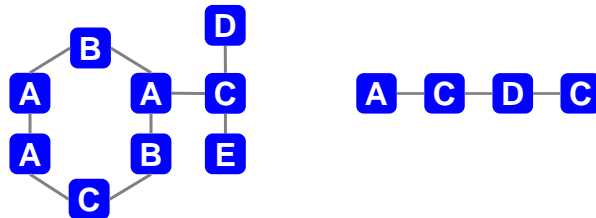
はじめに：構造データ解析のモチベーション

- 従来の機械学習の対象はベクトル形式で表現されていることが前提
 - その上での、学習アルゴリズムを議論
- これらの枠では捉えきれない「(グラフ)構造をもつデータ」が増加している
 - テキスト、HTML、XML、購買履歴、DNA配列、蛋白質立体構造、Web、遺伝子ネットワーク、社会ネットワーク、...

system system	parent parent	id id	parent_id parent_id	P. identifier P. identifier
system system	parent parent	1 1		P. identifier P. identifier
system system	parent parent	2 2	1	P. identifier P. identifier
system system	parent parent	3 3	1	P. identifier P. identifier
system system	parent parent	4 4	2	P. identifier P. identifier
system system	parent parent	5 5	2	P. identifier P. identifier
system system	parent parent	6 6	3	P. identifier P. identifier
system system	parent parent	7 7	3	P. identifier P. identifier
system system	parent parent	8 8	4	P. identifier P. identifier
system system	parent parent	9 9	4	P. identifier P. identifier
system system	parent parent	10 10	5	P. identifier P. identifier
system system	parent parent	11 11	5	P. identifier P. identifier
system system	parent parent	12 12	6	P. identifier P. identifier
system system	parent parent	13 13	6	P. identifier P. identifier
system system	parent parent	14 14	7	P. identifier P. identifier
system system	parent parent	15 15	7	P. identifier P. identifier
system system	parent parent	16 16	8	P. identifier P. identifier
system system	parent parent	17 17	8	P. identifier P. identifier
system system	parent parent	18 18	9	P. identifier P. identifier
system system	parent parent	19 19	9	P. identifier P. identifier
system system	parent parent	20 20	10	P. identifier P. identifier
system system	parent parent	21 21	10	P. identifier P. identifier
system system	parent parent	22 22	11	P. identifier P. identifier
system system	parent parent	23 23	11	P. identifier P. identifier
system system	parent parent	24 24	12	P. identifier P. identifier
system system	parent parent	25 25	12	P. identifier P. identifier
system system	parent parent	26 26	13	P. identifier P. identifier
system system	parent parent	27 27	13	P. identifier P. identifier
system system	parent parent	28 28	14	P. identifier P. identifier
system system	parent parent	29 29	14	P. identifier P. identifier
system system	parent parent	30 30	15	P. identifier P. identifier
system system	parent parent	31 31	15	P. identifier P. identifier
system system	parent parent	32 32	16	P. identifier P. identifier
system system	parent parent	33 33	16	P. identifier P. identifier
system system	parent parent	34 34	17	P. identifier P. identifier
system system	parent parent	35 35	17	P. identifier P. identifier
system system	parent parent	36 36	18	P. identifier P. identifier
system system	parent parent	37 37	18	P. identifier P. identifier
system system	parent parent	38 38	19	P. identifier P. identifier
system system	parent parent	39 39	19	P. identifier P. identifier
system system	parent parent	40 40	20	P. identifier P. identifier
system system	parent parent	41 41	20	P. identifier P. identifier
system system	parent parent	42 42	21	P. identifier P. identifier
system system	parent parent	43 43	21	P. identifier P. identifier
system system	parent parent	44 44	22	P. identifier P. identifier
system system	parent parent	45 45	22	P. identifier P. identifier
system system	parent parent	46 46	23	P. identifier P. identifier
system system	parent parent	47 47	23	P. identifier P. identifier
system system	parent parent	48 48	24	P. identifier P. identifier
system system	parent parent	49 49	24	P. identifier P. identifier
system system	parent parent	50 50	25	P. identifier P. identifier
system system	parent parent	51 51	25	P. identifier P. identifier
system system	parent parent	52 52	26	P. identifier P. identifier
system system	parent parent	53 53	26	P. identifier P. identifier
system system	parent parent	54 54	27	P. identifier P. identifier
system system	parent parent	55 55	27	P. identifier P. identifier
system system	parent parent	56 56	28	P. identifier P. identifier
system system	parent parent	57 57	28	P. identifier P. identifier
system system	parent parent	58 58	29	P. identifier P. identifier
system system	parent parent	59 59	29	P. identifier P. identifier
system system	parent parent	60 60	30	P. identifier P. identifier
system system	parent parent	61 61	30	P. identifier P. identifier
system system	parent parent	62 62	31	P. identifier P. identifier
system system	parent parent	63 63	31	P. identifier P. identifier
system system	parent parent	64 64	32	P. identifier P. identifier
system system	parent parent	65 65	32	P. identifier P. identifier
system system	parent parent	66 66	33	P. identifier P. identifier
system system	parent parent	67 67	33	P. identifier P. identifier
system system	parent parent	68 68	34	P. identifier P. identifier
system system	parent parent	69 69	34	P. identifier P. identifier
system system	parent parent	70 70	35	P. identifier P. identifier
system system	parent parent	71 71	35	P. identifier P. identifier
system system	parent parent	72 72	36	P. identifier P. identifier
system system	parent parent	73 73	36	P. identifier P. identifier
system system	parent parent	74 74	37	P. identifier P. identifier
system system	parent parent	75 75	37	P. identifier P. identifier
system system	parent parent	76 76	38	P. identifier P. identifier
system system	parent parent	77 77	38	P. identifier P. identifier
system system	parent parent	78 78	39	P. identifier P. identifier
system system	parent parent	79 79	39	P. identifier P. identifier
system system	parent parent	80 80	40	P. identifier P. identifier
system system	parent parent	81 81	40	P. identifier P. identifier
system system	parent parent	82 82	41	P. identifier P. identifier
system system	parent parent	83 83	41	P. identifier P. identifier
system system	parent parent	84 84	42	P. identifier P. identifier
system system	parent parent	85 85	42	P. identifier P. identifier
system system	parent parent	86 86	43	P. identifier P. identifier
system system	parent parent	87 87	43	P. identifier P. identifier
system system	parent parent	88 88	44	P. identifier P. identifier
system system	parent parent	89 89	44	P. identifier P. identifier
system system	parent parent	90 90	45	P. identifier P. identifier
system system	parent parent	91 91	45	P. identifier P. identifier
system system	parent parent	92 92	46	P. identifier P. identifier
system system	parent parent	93 93	46	P. identifier P. identifier
system system	parent parent	94 94	47	P. identifier P. identifier
system system	parent parent	95 95	47	P. identifier P. identifier
system system	parent parent	96 96	48	P. identifier P. identifier
system system	parent parent	97 97	48	P. identifier P. identifier
system system	parent parent	98 98	49	P. identifier P. identifier
system system	parent parent	99 99	49	P. identifier P. identifier
system system	parent parent	100 100	50	P. identifier P. identifier
system system	parent parent	101 101	50	P. identifier P. identifier
system system	parent parent	102 102	51	P. identifier P. identifier
system system	parent parent	103 103	51	P. identifier P. identifier
system system	parent parent	104 104	52	P. identifier P. identifier
system system	parent parent	105 105	52	P. identifier P. identifier
system system	parent parent	106 106	53	P. identifier P. identifier
system system	parent parent	107 107	53	P. identifier P. identifier
system system	parent parent	108 108	54	P. identifier P. identifier
system system	parent parent	109 109	54	P. identifier P. identifier
system system	parent parent	110 110	55	P. identifier P. identifier
system system	parent parent	111 111	55	P. identifier P. identifier
system system	parent parent	112 112	56	P. identifier P. identifier
system system	parent parent	113 113	56	P. identifier P. identifier
system system	parent parent	114 114	57	P. identifier P. identifier
system system	parent parent	115 115	57	P. identifier P. identifier
system system	parent parent	116 116	58	P. identifier P. identifier
system system	parent parent	117 117	58	P. identifier P. identifier
system system	parent parent	118 118	59	P. identifier P. identifier
system system	parent parent	119 119	59	P. identifier P. identifier
system system	parent parent	120 120	60	P. identifier P. identifier
system system	parent parent	121 121	60	P. identifier P. identifier
system system	parent parent	122 122	61	P. identifier P. identifier
system system	parent parent	123 123	61	P. identifier P. identifier
system system	parent parent	124 124	62	P. identifier P. identifier
system system	parent parent	125 125	62	P. identifier P. identifier
system system	parent parent	126 126	63	P. identifier P. identifier
system system	parent parent	127 127	63	P. identifier P. identifier
system system	parent parent	128 128	64	P. identifier P. identifier
system system	parent parent	129 129	64	P. identifier P. identifier
system system	parent parent	130 130	65	P. identifier P. identifier
system system	parent parent	131 131	65	P. identifier P. identifier
system system	parent parent	132 132	66	P. identifier P. identifier
system system	parent parent	133 133	66	P. identifier P. identifier
system system	parent parent	134 134	67	P. identifier P. identifier
system system	parent parent	135 135	67	P. identifier P. identifier
system system	parent parent	136 136	68	P. identifier P. identifier
system system	parent parent	137 137	68	P. identifier P. identifier
system system	parent parent	138 138	69	P. identifier P. identifier
system system	parent parent	139 139	69	P. identifier P. identifier
system system	parent parent	140 140	70	P. identifier P. identifier
system system	parent parent	141 141	70	P. identifier P. identifier
system system	parent parent	142 142	71	P. identifier P. identifier
system system	parent parent	143 143	71	P. identifier P. identifier
system system	parent parent	144 144	72	P. identifier P. identifier
system system	parent parent	145 145	72	P. identifier P. identifier
system system	parent parent	146 146	73	P. identifier P. identifier
system system	parent parent	147 147	73	P. identifier P. identifier
system system	parent parent	148 148	74	P. identifier P. identifier
system system	parent parent	149 149	74	P. identifier P. identifier
system system	parent parent	150 150	75	P. identifier P. identifier
system system	parent parent	151 151	75	P. identifier P. identifier
system system	parent parent	152 152	76	P. identifier P. identifier
system system	parent parent	153 153	76	P. identifier P. identifier
system system	parent parent	154 154	77	P. identifier P. identifier
system system	parent parent	155 155	77	P. identifier P. identifier
system system	parent parent	156 156	78	P. identifier P. identifier
system system	parent parent	157 157	78	P. identifier P. identifier
system system	parent parent	158 158	79	P. identifier P. identifier
system system	parent parent	159 159	79	P. identifier P. identifier
system system	parent parent	160 160	80	P. identifier P. identifier
system system	parent parent	161 161	80	P. identifier P. identifier
system system	parent parent	162 162	81	P. identifier P. identifier
system system	parent parent	163 163	81	P. identifier P. identifier
system system	parent parent	164 164	82	P. identifier P. identifier
system system	parent parent	165 165	82	P. identifier P. identifier
system system	parent parent	166 166	83	P. identifier P. identifier
system system	parent parent	167 167	83	P. identifier P. identifier
system system	parent parent	168 168	84	P. identifier P. identifier
system system	parent parent	169 169	84	P. identifier P. identifier
system system	parent parent	170 170	85	P. identifier P. identifier
system system	parent parent	171 171	85	P. identifier P. identifier
system system	parent parent	172 172	86	P. identifier P. identifier
system system	parent parent	173 173	86	P. identifier P. identifier
system system	parent parent	174 174	87	P. identifier P. identifier
system system	parent parent	175 175	87	P. identifier P. identifier
system system	parent parent	176 176	88	P. identifier P. identifier
system system	parent parent	177 177	88	P. identifier P. identifier
system system	parent parent	178 178	89	P. identifier P. identifier
system system	parent parent	179 179	89	P. identifier P. identifier
system system	parent parent	180 180	90	P. identifier P. identifier
system system	parent parent	181 181	90	P. identifier P. identifier
system system	parent parent	182 182	91	P. identifier P. identifier
system system	parent parent	183 183	91	P. identifier P. identifier
system system	parent parent	184 184	92	P. identifier P. identifier
system system	parent parent	185 185	92	P. identifier P. identifier
system system	parent parent	186 186	93	P. identifier P. identifier
system system	parent parent	187 187	93	P. identifier P. identifier
system system	parent parent	188 188	94	P. identifier P. identifier
system system	parent parent	189 189	94	P. identifier P. identifier
system system	parent parent	190 190	95	P. identifier P. identifier
system system	parent parent	191 191	95	P. identifier P. identifier
system system	parent parent	192 192	96	P. identifier P. identifier
system system	parent parent	193 193	96	P. identifier P. identifier
system system	parent parent	194 194	97	P. identifier P. identifier
system system	parent parent	195 195	97	P. identifier P. identifier
system system	parent parent	196 196	98	P. identifier P. identifier
system system	parent parent	197 197	98	P. identifier P. identifier
system system	parent parent	198 198	99	P. identifier P. identifier
system system	parent parent	199 199	99	P. identifier P. identifier
system system	parent parent	200 200	100	P. identifier P. identifier
system system	parent parent	201 201	100	P. identifier P. identifier
system system	parent parent	202 202	101	P. identifier P. identifier
system system	parent parent	203 203	101	P. identifier P. identifier
system system	parent parent	204 204	102	P. identifier P. identifier
system system	parent parent	205 205	102	P. identifier P. identifier
system system	parent parent	206 206	103	P. identifier P. identifier
system system	parent parent	207 207	103	P. identifier P. identifier
system system	parent parent	208 208	104	P. identifier P. identifier
system system	parent parent	209 209	104	P. identifier P. identifier
system system	parent parent	210 210	105	P. identifier P. identifier
system system	parent parent	211 211	105	P. identifier P. identifier
system system	parent parent	212 212	106	P. identifier P. identifier
system system	parent parent	213 213	106	P. identifier P. identifier
system system	parent parent	214 214	107	P. identifier P. identifier
system system	parent parent	215 215	107	P. identifier P. identifier
system system	parent parent	216 216	108	P. identifier P. identifier
system system	parent parent	217 217	108	P. identifier P. identifier
system system	parent parent	218 218	109	P. identifier P. identifier
system system	parent parent	219 219	109	P. identifier P. identifier
system system	parent parent	220 220	110	P. identifier P. identifier
system system	parent parent	221 221	110	P. identifier P. identifier
system system	parent parent	222 222	111	P. identifier P. identifier
system system	parent parent	223 223	111	P. identifier P. identifier
system system	parent parent	224 224	112	P. identifier P. identifier
system system	parent parent	225 225	112	P. identifier P. identifier
system system	parent parent	226 226	113	P. identifier P. identifier
system system	parent parent	227 227	113	P. identifier P. identifier
system system	parent parent	228 228	114	P. identifier P. identifier
system system	parent parent	229 229	114	P. identifier P. identifier
system system	parent parent	230 230	115	P. identifier P. identifier
system system	parent parent	231 231	115	P. identifier P. identifier
system system	parent parent	232 232	116	P. identifier P. identifier

本講演の内容：カーネル法による構造データの解析手法

- 1) カーネル法とは
- 2) 構造カーネル法 – 畳み込みカーネルを中心に
- 3) 分類問題から、構造マッピング問題へ



1) カーネル法とは？

- カーネル法の例
 - 2クラスの分類学習問題を題材に紹介
 - パーセプトロンを使った説明
- カーネル法の一般的な概念

題材: 2クラスの分類学習問題

■ 目的: 写像 $h: X \rightarrow Y$ を求める

- X : 全ての解析対象の集合
 - 例: 人間全体の集合
- $Y = \{+1, -1\}$: 解析対象が属するクラスの集合
 - 例: 男 / 女

■ 手がかり

- N 個の訓練データ: $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(N)}, y^{(N)})$
 - ただし, $(x^{(i)}, y^{(i)}) \in X \times Y$

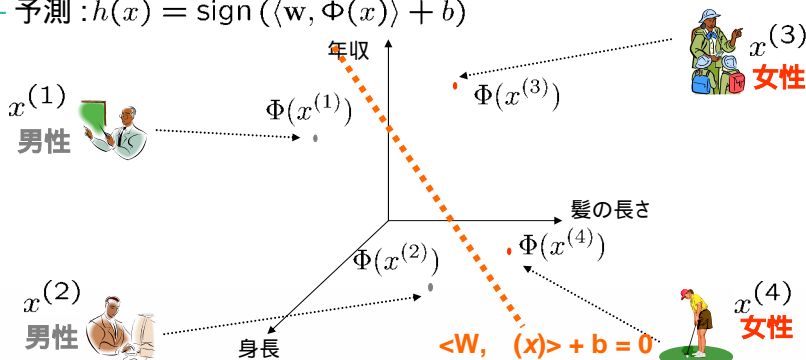


© 2005 IBM Corporation

2クラスの分類学習問題への一般的アプローチ

- 解析対象 $x \in X$ を、特徴空間中の点(=特徴ベクトル) $\Phi(x)$ として表現
- それぞれのクラスに属する点集合を分類する面を求める

- 超平面: $\langle w, \Phi(x) \rangle + b = 0$
 - (w, b) : 重みベクトル
- 予測: $h(x) = \text{sign}(\langle w, \Phi(x) \rangle + b)$



パーセプトロン

■ 2クラス分類の逐次学習アルゴリズム

– 訓練データをひとつずつ処理しながら学習

– 1) 訓練データ $x^{(i)}$ に対する予測を行う

$$h(x^{(i)}) = \text{sign}(\langle \mathbf{w}, \Phi(x^{(i)}) \rangle + b)$$

– 2) 予測が外れたとき ($h(x^{(i)}) \neq y^{(i)}$) のみ、重みベクトルを更新

- $\mathbf{w} \leftarrow \mathbf{w} + y^{(i)} \Phi(x^{(i)})$

$$b \leftarrow b + y^{(i)} R^2$$

重みベクトルに
特徴ベクトルを足す / 引く
($y^{(i)} \in \{+1, -1\}$)

パーセプトロンからカーネル法へ

■ 学習 = 重みベクトル \mathbf{w} に特徴ベクトルを足す / 引く

$$\mathbf{w} \leftarrow \mathbf{w} + y^{(i)} \Phi(x^{(i)})$$

■ つまり、 \mathbf{w} は訓練データの特徴ベクトルの線形和で表せる

$$\mathbf{w} = \sum_{i=1}^N \alpha_i \Phi(x^{(i)})$$

– α_i は新たな重みパラメータ

■ カーネル法としての、パーセプトロン

– 代入による分類器の書き換え

$$h(x) = \text{sign} \left(\sum_{i=1}^N \alpha_i \underbrace{\langle \Phi(x^{(i)}), \Phi(x) \rangle}_{\text{内積による}} + b \right)$$

– カーネル関数: 内積の置き換え **データアクセス**

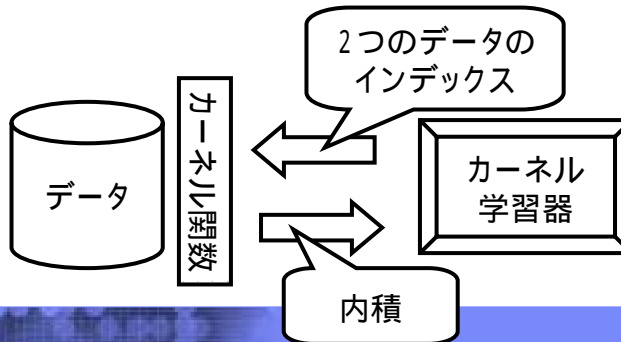
$$K(x, x') = \langle \Phi(x), \Phi(x') \rangle$$

カーネル法とは

- **カーネル関数 (= 内積) によってのみデータアクセスを行う学習器**

$$K(x, x') = \langle \Phi(x), \Phi(x') \rangle \quad (\text{カーネル関数})$$

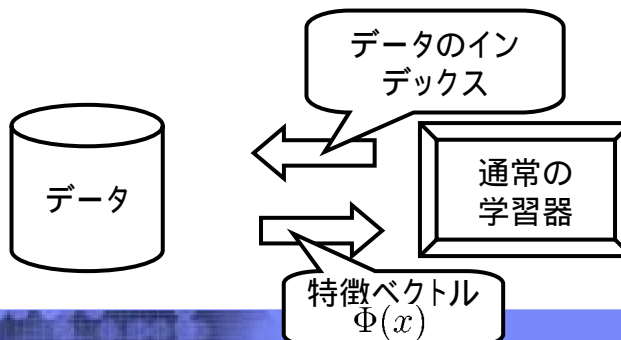
$$h(x) = \text{sign} \left(\sum_{i=1}^N \alpha_i K(x^{(i)}, x) \right) \quad (\text{カーネル分類器})$$



© 2005 IBM Corporation

カーネル法のポイント

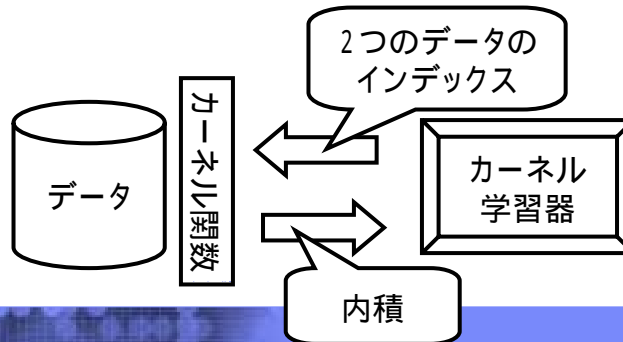
- **特徴ベクトル $\Phi(x)$ が陽に現れない**
 - (カーネル関数はさておき) データアクセス部分が次元に依存しない
- **x や x' が何であれ、適当に類似度 $K(x, x')$ を定義すればそれがカーネル関数として使える**
 - 特徴ベクトルを陽に設計しなくてもよい



© 2005 IBM Corporation

カーネル法のポイント

- 特徴ベクトル $\Phi(x)$ が陽に現れない
 - (カーネル関数はさておき) データアクセス部分が次元に依存しない
- x や x' が何であれ、適当に類似度 $K(x, x')$ を定義すればそれがカーネル関数として使える
 - 特徴ベクトルを陽に設計しなくてもよい



© 2005 IBM Corporation

2) 構造カーネル法

- 構造をもったデータ
 - 内部構造と外部構造
- 構造データに対するカーネル
 - 畳み込みカーネルの概念
 - グラフカーネルの設計
 - 様々な構造に対する畳み込みカーネル

© 2005 IBM Corporation

構造をもったデータ

■ 外部構造: データ間にグラフ構造がある

- Web
- 社会ネットワーク
- 遺伝子 / 蛋白質ネットワーク

■ 本日のお話

■ 内部構造: データ内にグラフ構造がある

- HTML、XML
- DNA
- 化合物

データ間の関係

データ

データの
構成要素

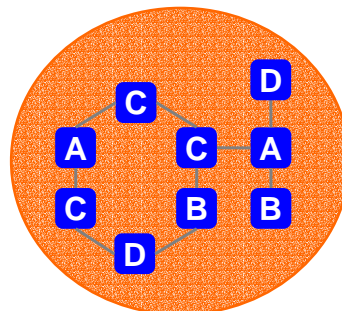
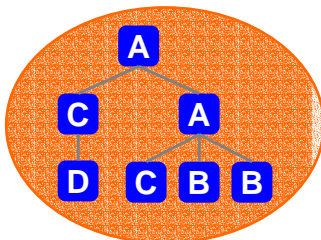
要素間の関係

© 2005 IBM Corporation

当面の目標: グラフとグラフの間のカーネル関数をつくる

■ 仕様

- グラフとグラフの類似度をうまく表している
- それなりの時間で計算できる

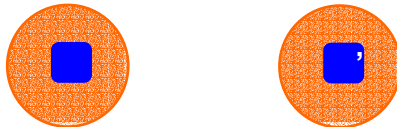


簡単なケース1: 単一ノード同士のカーネル関数

- ラベルを比較して、同じなら1、異なるなら0

$$K_{\Sigma}(\sigma, \sigma') = \begin{cases} 1 & (\sigma = \sigma') \\ 0 & (\sigma \neq \sigma') \end{cases}$$

- あるいは、何らかの類似度を定義
 - 原子の性質、大文字 / 小文字、...



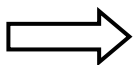
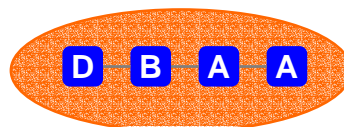
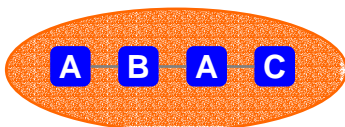
簡単なケース2: 長さ n の配列同士のカーネル関数

- ノード同士のカーネル関数の積

$$K_S(s, s') = \prod_{i=1}^n K_{\Sigma}(\sigma(s_i), \sigma(s'_i))$$

$$s = (s_1, s_2, \dots, s_n)$$

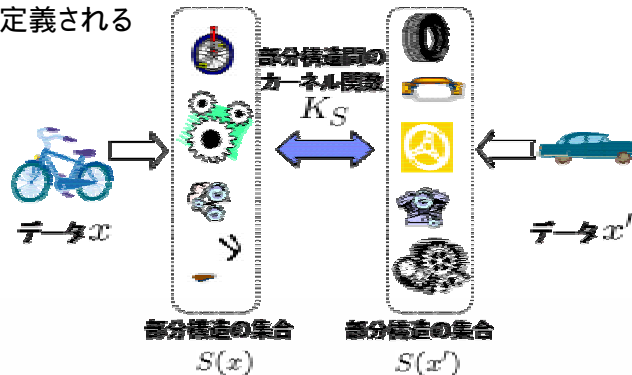
$$s' = (s'_1, s'_2, \dots, s'_n)$$



この調子で拡張していても
少しでも形が異なった時点でアウト

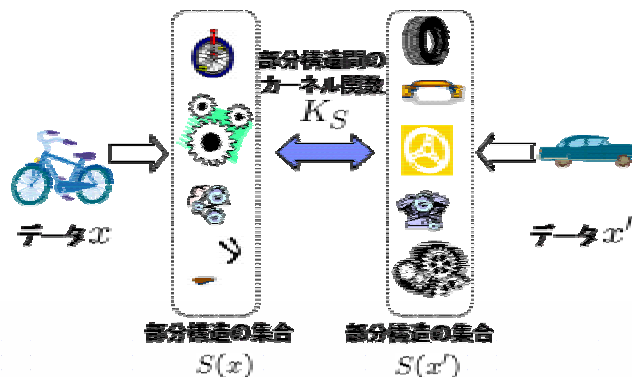
畳み込みカーネル: 構造カーネル設計の一般的な枠組み

- **構造データの特徴は、その部分構造が担っている**
 - 例: 自動車の特徴は、部品の特徴が担っている
- **2つの構造の間のカーネル関数は、部分構造間のカーネル関数によって再帰的に定義される**
 - 例: 自転車と自動車のカーネル関数は、それぞれの部品同士のカーネル関数によって定義される



畳み込みカーネルの定義

- **定義:**
$$K(x, x') = \sum_{s \in S(x)} \sum_{s' \in S(x')} K_S(s, s')$$
 - $S(x)$: x の部分構造の集合
 - K_S : 部分構造間のカーネル関数



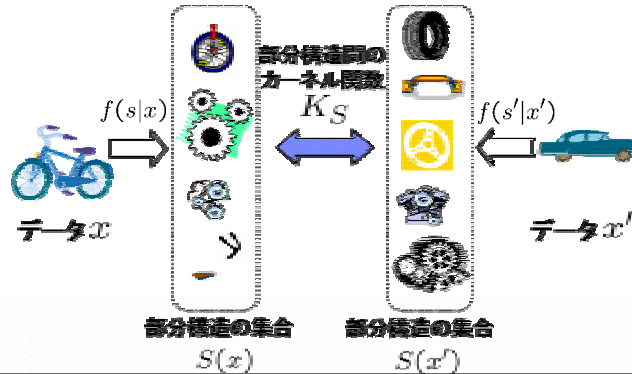
(重み付き) 畳み込みカーネルの定義

■ **定義:**
$$K(x, x') = \sum_{s \in S(x)} \sum_{s' \in S(x')} f(s|x) f(s'|x') K_S(s, s')$$

- $S(x)$: x の部分構造の集合
- K_S : 部分構造間のカーネル関数
- $f(s|x)$: x の部分構造 $s \in S(x)$ の重み

これらを定義

↓
(再帰)計算

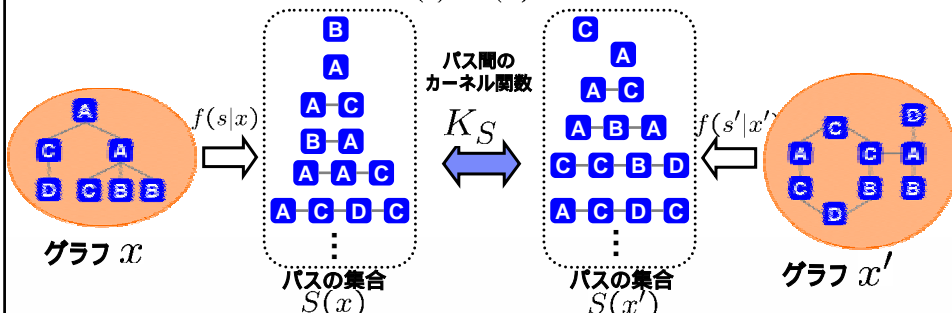


グラフ・カーネルの定義

- **部分構造の集合 $S(x)$ として何を使うか？**

- 部分グラフ？ **ダメ** (計算がNP-hard)
- グラフ上のランダム・ウォークによって生成されるパスを使う！
 - K_S は同じ長さの配列同士のカーネル
 - $f(s|x) = \lambda^{|s|}$ はパス s の長さによって減衰 ($0 < \lambda < 1$)

$$K(v, v') = \sum_{s \in S(x)} \sum_{s' \in S(x')} \lambda^{|s|} \lambda^{|s'|} K_S(s, s')$$



$$K(v, v') = \sum_{s \in S(x)} \sum_{s' \in S(x')} \lambda^{|s|} \lambda^{|s'|} K_S(s, s')$$

グラフ・カーネルの計算

- グラフ上のランダム・ウォークによって生成されるパスは無限個ありうるため、ナイーブな計算はできない

- 再帰計算

– $S_v(x)$: ノード v で終わるパスの集合 ($S(x) = \cup_{v \in V} S_v(x)$)

– ノード対 (v, v') で終わるパスのみに注目したときのカーネル K_V の

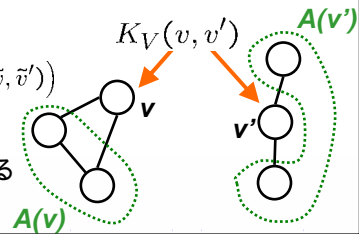
和に分解
$$K_V(v, v') = \sum_{s \in S_v(x)} \sum_{s' \in S_{v'}(x')} \lambda^{|s|} \lambda^{|s'|} K_S(s, s')$$

$$K(x, x') = \sum_{v \in V} \sum_{v' \in V'} K_V(v, v')$$

– K_V が再帰的に書ける!

$$K_V(v, v') = \lambda^2 K_\Sigma(v, v') \left(1 + \sum_{\tilde{v} \in A(v)} \sum_{\tilde{v}' \in A(v')} \lambda^2 K_V(\tilde{v}, \tilde{v}') \right)$$

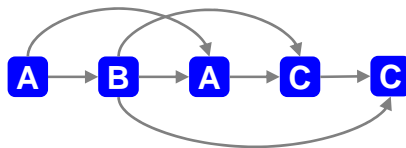
- 連立方程式を解けばカーネルが計算できる (多項式時間)



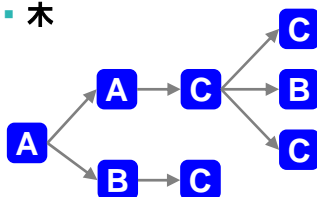
特殊な場合

- 非循環有向グラフ (DAG)

– 再帰式がループしないので、動的計画法で、より速く解ける



- 木



- 配列



様々な畳み込みカーネル

対象 部分構造	配列	順序木	木	DAG	グラフ
パス	(2)	(2)	(2)		(1)
部分グラフ			×	×	×

1 : 部分構造が同じノードを何度も使える

2 : $K_{\Sigma}(\sigma, \sigma') = \begin{cases} 1 & (\sigma = \sigma') \\ 0 & (\sigma \neq \sigma') \end{cases}$ のときには接尾辞木による線形時間解法

構造カーネルのまとめ

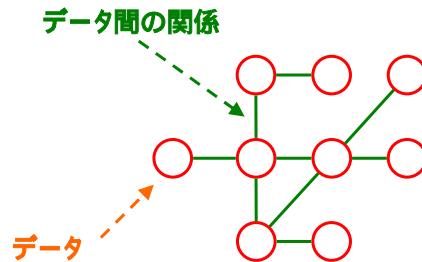
- データのもつ構造は外部構造と内部構造にわけられる
- 内部構造をもつデータのカーネル関数は、部分構造間のカーネル関数によって再帰的に定義される（畳み込みカーネル）
- 部分構造の数は通常、非常に多くなるので畳み込みカーネルのナイーブな計算は困難
- 畳み込みカーネルは、通常、再帰計算によって効率よく計算することができる

補足: 外部構造に対するカーネル関数

- 隣り合うデータ間のカーネル関数が与えられたとき、離れたデータ間のカーネル関数を定義する

— 例: 拡散カーネル

- 隣の隣は隣

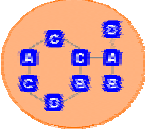
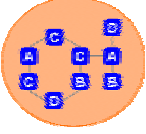
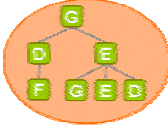


3) 構造マッピング問題

- 構造マッピング問題の定義
- 2つのアプローチ
- 隠れマルコフ・パーセプトロン
- 構造カーネルの適用と問題点
- 2段階学習による解決

構造マッピング問題

- 構造分類問題の、より一般的なケース

	X	Y
2クラス分類	 構造データ	$\{+1, -1\}$
構造マッピング	 構造データ	 構造データ

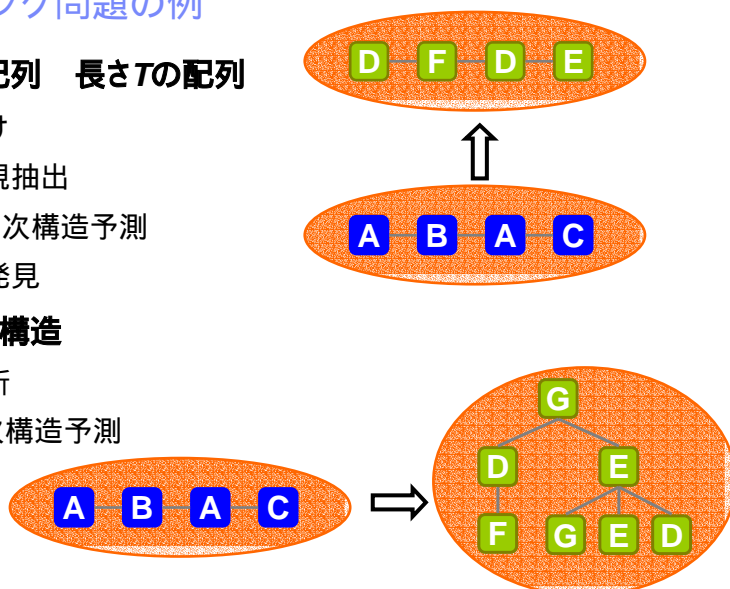
構造マッピング問題の例

- 長さ T の配列 長さ T の配列

- 品詞付け
- 固有表現抽出
- 蛋白質2次構造予測
- 遺伝子発見

- 配列 木構造

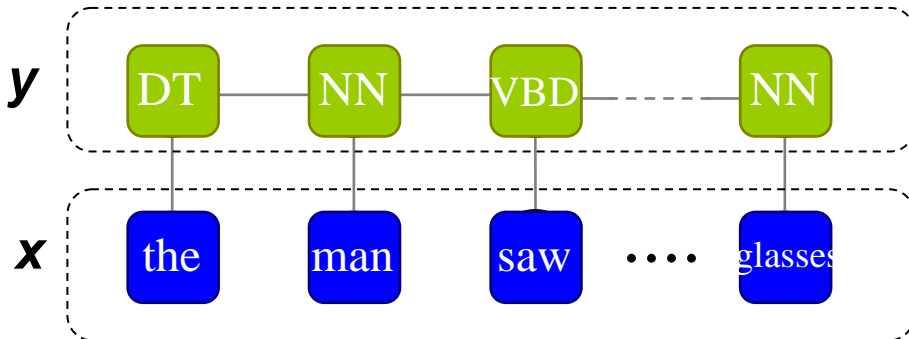
- 構文解析
- RNA2次構造予測



品詞付け

- 長さ T の配列 長さ T の配列のマッピング問題

- 構造は固定、ラベルだけ当てる



© 2005 IBM Corporation

構造マッピング問題への2つのアプローチ

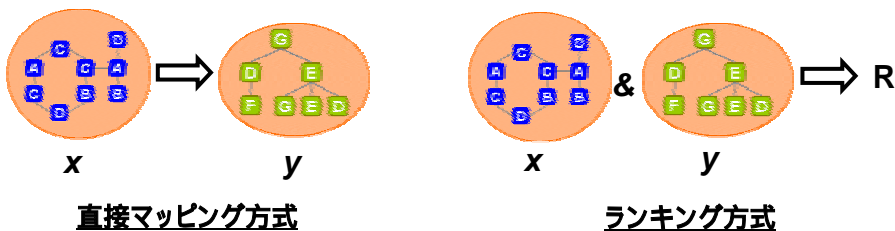
- 直接マッピング方式: x の特徴空間から y の特徴空間へのマッピングを直接求める

- $\Phi(y) = h(\Phi(x))$ を求め、 $y = \Phi^{-1}(\Phi(y))$

本日のお話

- ランキング方式: x と y を合わせた構造に対する良し悪しを評価する

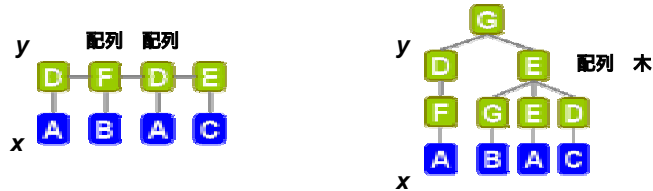
- x と y を合わせた構造の特徴ベクトル $\Phi(x, y)$ のランキング問題



© 2005 IBM Corporation

ランキング方式

- x と y を合わせた構造 に対する特徴ベクトル $\Phi(x, y)$ を評価



- 重み w による評価値が最大になる y を出力する

$$\hat{y} = \operatorname{argmax}_{y \in Y} \langle w, \Phi(x, y) \rangle$$



隠れマルコフ・パーセプトロン

- ランキング方式の構造マッピング学習アルゴリズム

- 逐次学習アルゴリズム

- 訓練データをひとつずつ処理しながら学習

- 1) 訓練データ $x^{(i)}$ に対する予測を行う

$$\hat{y}^{(i)} = \operatorname{argmax}_{y \in Y} \langle w, \Phi(x^{(i)}, y) \rangle$$

- 2) 予測が外れたとき ($\hat{y}^{(i)} \neq y^{(i)}$) のみ、重みベクトルを更新

- $w \leftarrow w + \Phi(x^{(i)}, y^{(i)}) - \Phi(x^{(i)}, \hat{y}^{(i)})$

- 全ての誤ったマッピング $y \neq \hat{y}^{(i)}$ に対し、

$$\langle w^*, \Phi(x^{(i)}, y^{(i)}) \rangle > \langle w^*, \Phi(x^{(i)}, \hat{y}^{(i)}) \rangle$$

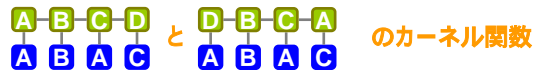
やはり、
重みベクトルに
特徴ベクトルを足す&引く

隠れマルコフ・カーネル・パーセプトロン

- パーセプトロンと同様、カーネル化できる

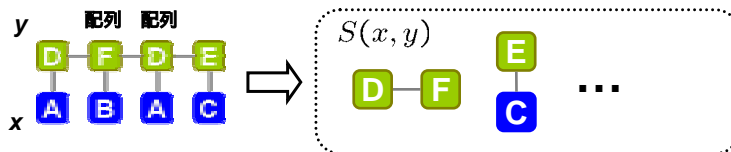
$$\begin{aligned}\hat{y}^{(i)} &= \operatorname{argmax}_{y \in Y} \langle \mathbf{w}, \Phi(x^{(i)}, y) \rangle \\ \hat{y}^{(i)} &= \operatorname{argmax}_{y \in Y} \sum_{j=1}^N \sum_{\tilde{y} \in Y} \alpha_j(\tilde{y}) \langle \Phi(x^{(j)}, \tilde{y}), \Phi(x^{(i)}, y) \rangle \\ &= \operatorname{argmax}_{y \in Y} \sum_{j=1}^N \sum_{\tilde{y} \in Y} \alpha_j(\tilde{y}) K((x^{(j)}, \tilde{y}), (x^{(i)}, y))\end{aligned}$$

マッピング同士のカーネル関数
(組み合わせた構造)



マッピング(組み合わせた構造)同士のカーネル関数

- 畳み込みカーネル $K((x, y), (x', y'))$ が使える
- 組み合わせた構造の部分構造 $S(x, y)$



— 注意: y の要素は、任意個は使えない

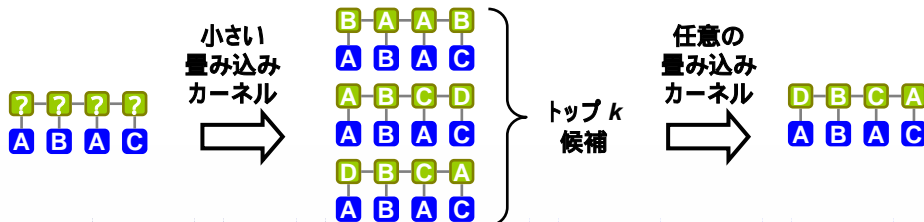
- argmax 操作が、部分構造の含む y 要素の数に対して指数的な計算量
- Viterbi アルゴリズム (動的計画法) と同じ理由



2段階学習: 再ランキングによる方法

- 小さい部分構造にもとづく隠れマルコフ・パーセプトロンによって、トップ k 候補を出力
 - ・ 但し、この中に正解が含まれるかは保証されない
- そのなかで一番よいものを、大きい部分構造を用いた隠れマルコフ・パーセプトロンで予測
 - あらかじめ候補が k 個に絞られているので、Viterbiの必要なし

$$\hat{y}^{(i)} = \operatorname{argmax}_{y \in \text{トップ } k \text{ 候補}} \langle \mathbf{w}, \Phi(x^{(i)}, y) \rangle$$



構造マッピングのまとめ

- 構造マッピング問題には、直接マッピング方式と、ランキング方式がある
- ランキング方式の一手法が隠れマルコフパーセプトロンであり、パーセプトロンと同様、カーネル化できる
- 構造マッピング問題においては、畳み込みカーネルで用いることのできる部分構造に制限がある
- そのため、あくまで任意サイズの部分構造を用いたい場合には、学習を2段階にするなどの工夫が必要

構造カーネルの課題

■ 高速化

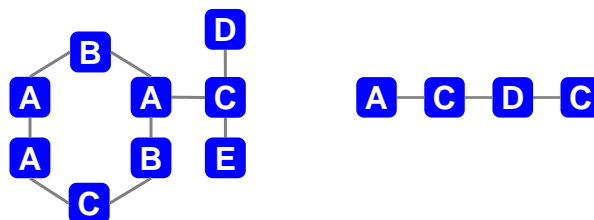
- 動的計画法などによって、効率的に計算できるといっても、遅い
 - より高速なアルゴリズム（線形時間、確率、近似）
- カーネル法の計算量は、訓練例の数に依存
 - 例の効率的な管理（変換、データ構造）

■ 特徴選択(属性選択)による精度向上

- カーネル関数によって特徴空間表現が隠蔽されているため、特徴選択が陽に行えない
 - 部分構造の重み $f(s|x)$ の学習が特徴選択に相当

おわり

- 1) カーネル法とは
- 2) 構造カーネル法 – 畳み込みカーネルを中心に
- 3) 分類問題から、構造マッピング問題へ



ありがとうございました

