

テーマ：構造データ解析のための機械学習手法

- 構造をもったデータを効率的に解析するための機械学習手法を研究しました
- 特にカーネル法とよばれる手法を用いた構造データの解析手法に貢献しました
- また、生体ネットワークなどのネットワーク構造の解析手法の研究を行いました

生命システム情報学講座 バイオ情報ネットワーク分野
指導教員：阿久津 達也 教授

鹿島 久嗣

構成

- 構造データの解析
 - 構造データとは
 - 構造データの種類
- 解析手法
 - カーネル法による構造データの解析手法
 - カーネル法（畳み込みカーネル）とは
 - 構造データに対するカーネル関数の設計
 - [研究成果1] ラベル付順序木カーネルについて
 - [研究成果2] グラフカーネルについて
 - 構造データカーネルの発展
 - [研究成果3] 構造ラベル付けについて
 - ネットワーク構造をもったデータの解析手法
 - リンク予測について
 - [研究成果4] ネットワーク構造の遷移モデルに基づくリンク予測について
- まとめ／発表文献

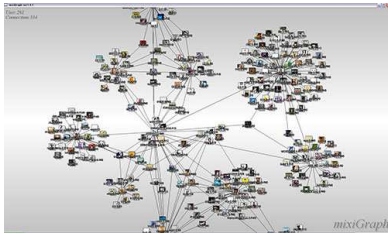
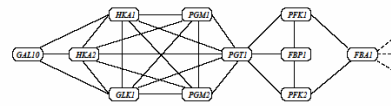
近年、構造をもったデータの解析の重要性が高まっています

▪ 近年、構造をもった電子的データが増加している

▶ たとえば、

- 配列データ: DNA、タンパク質、自然言語、
- 木構造データ: HTML/XML、RNA構造、構文解析木、系統樹、ディレクトリ
- グラフ構造データ: 化合物、WWW、SNS、生体ネットワーク

▪ これらの解析が必要!



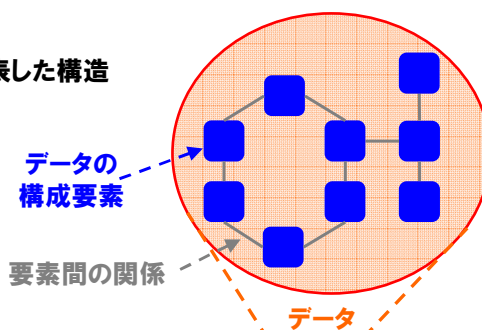
AGERTL	L	YVESIS	Y	VYLSGCT	F.	knowled
AGERTL	L	YVESIS	Y	VYLSGCT	F.	silivole
AGERTL	L	YVESIS	Y	VYLSGCT	F.	Salvador
AGERTL	L	YVESIS	Y	VYLSGCT	F.	Salvador-2
AGERTL	L	YVESIS	Y	VYLSGCT	F.	Africa
AGERTL	L	YVESIS	Y	VYLSGCT	F.	Thal-100
AGERTL	L	YVESIS	Y	VYLSGCT	F.	Thal-115
AGERTL	L	YVESIS	Y	VYLSGCT	F.	UK.Korea
AGERTL	L	YVESIS	Y	VYLSGCT	F.	Vietnam
AGERTL	L	YVESIS	Y	VYLSGCT	F.	Salvador-1
AGERTL	L	YVESIS	Y	VYLSGCT	F.	Salvador-2
AGERTL	L	YVESIS	Y	VYLSGCT	F.	Brazil-1
AGERTL	L	YVESIS	Y	VYLSGCT	F.	Brazil-2
AGERTL	L	YVESIS	Y	VYLSGCT	F.	Honduras-1
AGERTL	L	YVESIS	Y	VYLSGCT	F.	Honduras-2
AGERTL	L	YVESIS	Y	VYLSGCT	F.	Parana

「構造データ」と一口に言っても...

「構造」には内部構造と外部構造の2種類があり、扱いが異なります

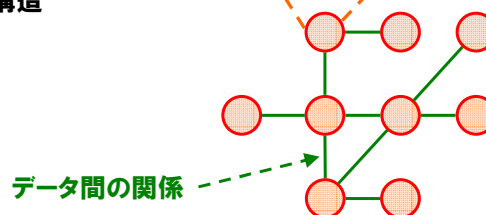
▪ 内部構造: データ内の要素の関連を表した構造

- ▶ HTML、XML
- ▶ DNA
- ▶ 化合物



▪ 外部構造: データ間の関連を表した構造

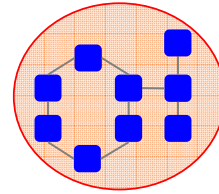
- ▶ Web
- ▶ 社会ネットワーク
- ▶ 遺伝子/蛋白質ネットワーク



私は、内部構造と外部構造のそれぞれの取り扱いに取り組み、いくつかの研究成果を残しました

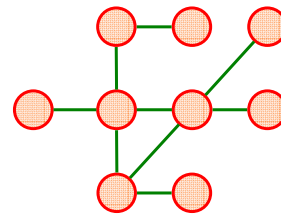
▪ 内部構造の扱い

- ▶ カーネル法に基づく解析手法
 - [研究成果 1] ラベル付順序木カーネル
 - [研究成果 2] グラフ・カーネル
 - [研究成果 3] ラベル付けカーネルマシン



▪ 外部構造の扱い

- ▶ リンク予測問題への確率的アプローチ
 - [研究成果 4] ネットワーク構造におけるリンク予測



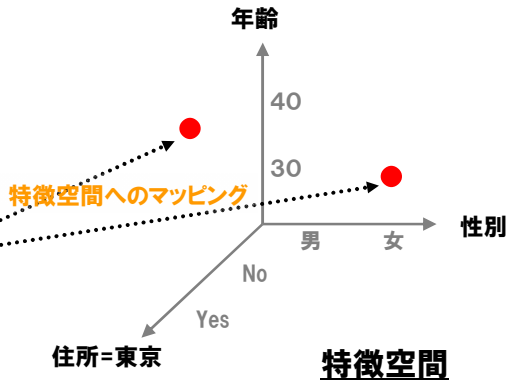
内部構造の扱いについての研究成果

背景：従来の機械学習手法ではベクトル型のデータを前提としているため、構造をもったデータをうまく扱うことが出来ません

- 従来の機械学習手法では、データが特徴空間中のベクトルとして表されていることを前提とする
- 構造をもったデータでは、特徴空間の構成が自明でない

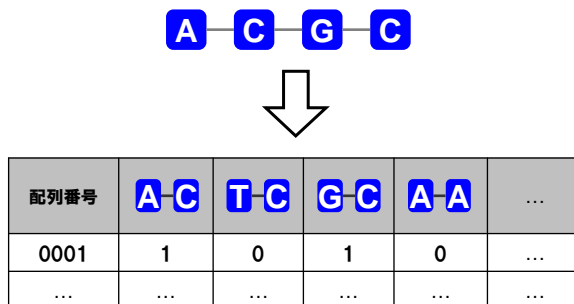
従来：ベクトル型のデータ

顧客番号	顧客氏名	年齢	性別	住所	...
0001	〇〇	40代	男性	東京都	...
0002	××	30代	女性	大阪府	...



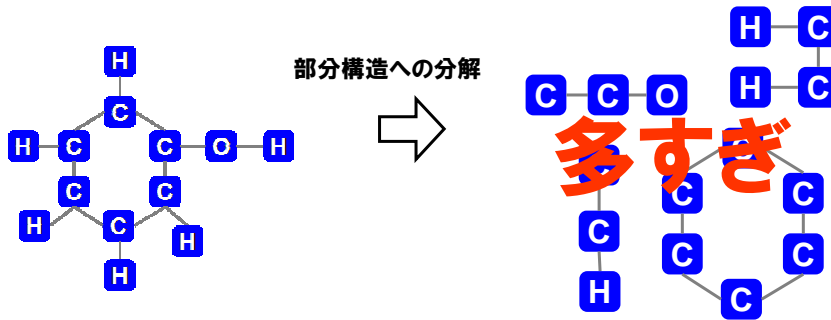
構造データを用いるための1つのアプローチは「部分構造」を用いて特徴空間を構成することです

- 構造データの解析では、「部分構造」が構造の性質を担っていると考える
 - ↳ 配列データの性質は、部分配列が担う(=マルコフモデル)
- たとえば、各部分構造の有無や出現回数を使って特徴空間を構成し、ベクトル表現する



課題：あわよくば「すべての」部分構造の情報を使いたいが、
ナイーブに数え上げるのは計算上のムリがあります

- すべての部分構造をテーブル要素の候補にしたい
- グラフには、指数個のサブグラフが含まれるので、すべてを数え上げている場所と時間はない ⇒ ここをいかに解決するかがポイント
- 考えうるアプローチ：
 - ▶ 部分構造のサイズを限定する ← マルコフモデル
 - ▶ 何らかの条件をつけて限られた重要なものだけを数え上げる？ ← バタンマイニング
 - ▶ そもそもベクトル表現に持ち込まない？ ← カーネル法（こちらを採択）



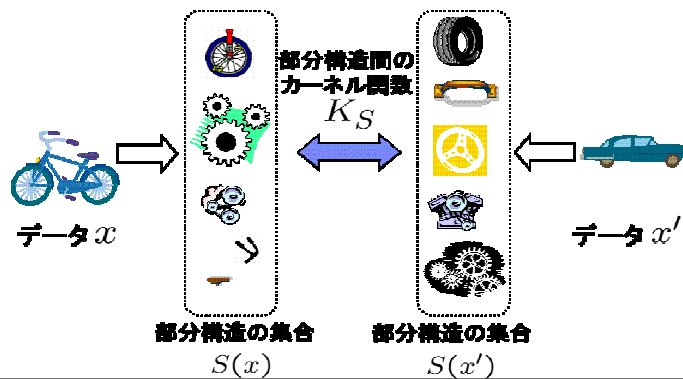
私は「畳み込みカーネル法」と呼ばれるすべての部分構造にもとづく
アプローチに着目しました

- カーネル法：カーネル関数(=2つのデータの類似度)に基づく機械学習法
 - ▶ 良いカーネル関数を定義できれば、あとはおまかせで動く → ポイントはカーネル関数の設計
- 畳み込みカーネル：2つの構造データのカーネル関数(=2つのデータの類似度)を、
共通にもつ部分構造を用いて定義する方法
 - ▶ 対象とする構造データに対して、部分構造の定義と、数え方の定義をする
 - ▶ カーネル関数の計算は、共通部分構造の数え上げの問題になる



畳み込みカーネルは、
部分構造に基づく構造カーネル設計の一般的な枠組みです

- 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

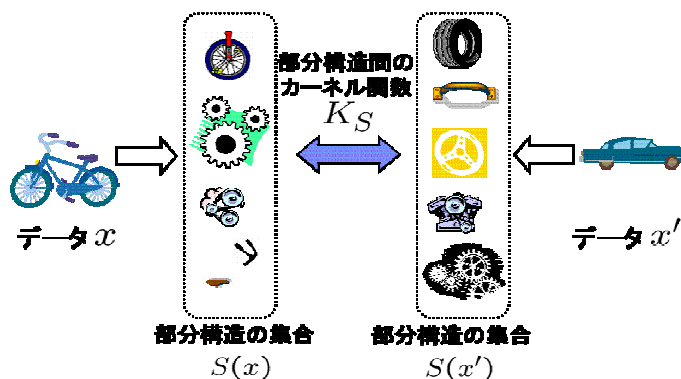


畳み込みカーネルの定義: 部分構造間のカーネルの和

▪ 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

これらを定義
↓
(再帰)計算



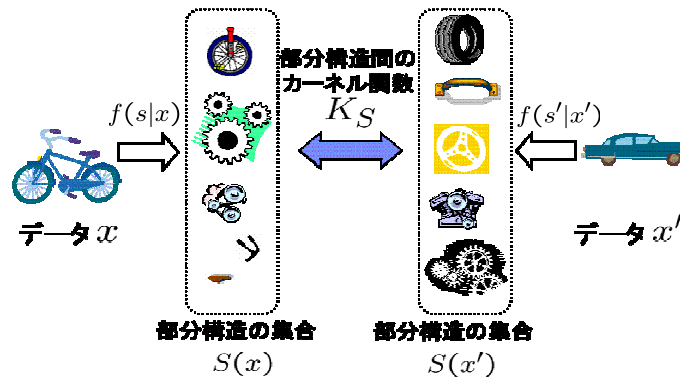
重付き畳み込みカーネルの定義：部分構造カーネルの重みつき和

▪ 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

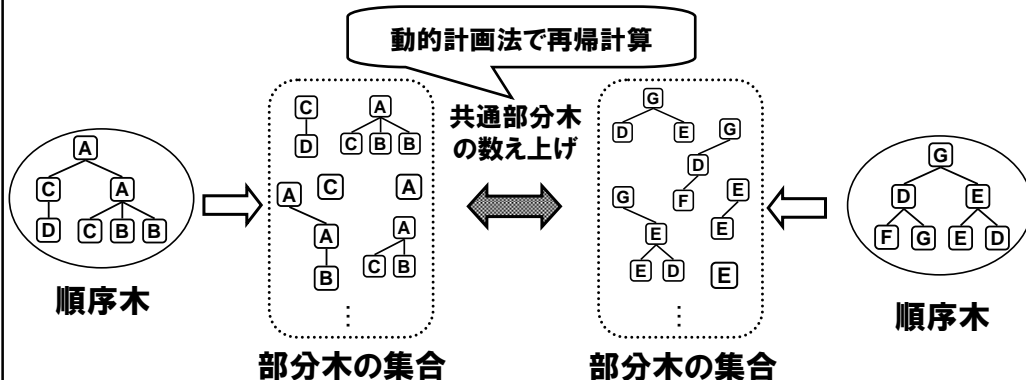
これらを定義

(再帰)計算



[研究成果1] 順序木に対するカーネル関数の設計を行いました

- 順序木カーネルでは、カーネル関数を
 - ▶ 部分構造 = 部分木
 - ▶ 数え方 = 2つの順序木に共通の部分木の個数
 によって定義する
- 難しいところ：部分木は指数個ありうる
 - ← 解決法：数え上げの計算を動的計画法によって再帰計算することで計算可能になる



H. Kashima and T. Koyanagi: Kernels for Semi-Structured Data, In Proc. 19th International Conference on Machine Learning (ICML), 2002
 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

順序木カーネルの計算は、再帰によって効率的に行えます

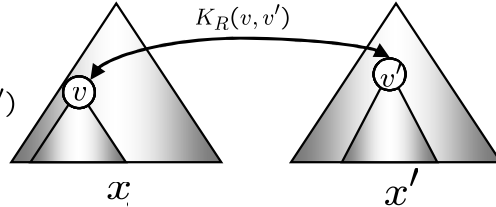
- 部分構造の集合 $S(x)$ として部分木をつかう

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

- カーネル関数 K は v と v' を根に持つ部分木に限ったカーネル K_R の和としてかける

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

$$K_R(v, v') = \sum_{s \in S_v(x)} \sum_{s' \in S_{v'}(x')} K_S(s, s')$$



- K_R は「 (v, v') についてボトムアップ」と「 (v, v') の子について左から右」の2重ループの動的計画法によって $O(|V||V'|)$ で再帰計算

ノードペア
(v, v')
について

(v, v')の
子同士
について

$$K^R(v, v') = I(\ell(v) = \ell(v')) \cdot \bar{K}_{v, v'}^R(\#ch(v), \#ch(v'))$$

$$\bar{K}_{v, v'}^R(i, j) = \bar{K}_{v, v'}^R(i-1, j) + \bar{K}_{v, v'}^R(i, j-1) - \bar{K}_{v, v'}^R(i-1, j-1) + \bar{K}_{v, v'}^R(i-1, j-1) \cdot K^R(ch(v, i), ch(v', j))$$

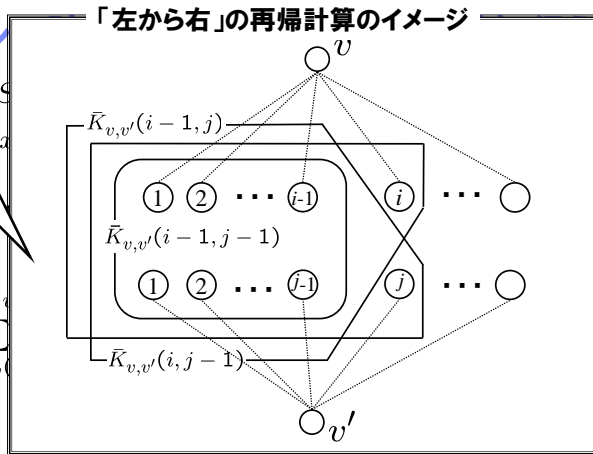
順序木カーネルの計算は、再帰によって効率的に行えます

- 部分構造の集合 $S(x)$ として部分木をつかう

(v, v') の子同士の
マッチングを
全て数え上げる

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

$$K_R(v, v') = \sum_{s \in S_v(x)} \sum_{s' \in S_{v'}(x')} K_S(s, s')$$



てかける

x'

- K_R は「 (v, v') についてボトムアップ」と「 (v, v') の子について左から右」の2重ループの動的計画法によって $O(|V||V'|)$ で再帰計算

ノードペア
(v, v')
について

(v, v')の
子同士
について

$$K^R(v, v') = I(\ell(v) = \ell(v')) \cdot \bar{K}_{v, v'}^R(\#ch(v), \#ch(v'))$$

$$\bar{K}_{v, v'}^R(i, j) = \bar{K}_{v, v'}^R(i-1, j) + \bar{K}_{v, v'}^R(i, j-1) - \bar{K}_{v, v'}^R(i-1, j-1) + \bar{K}_{v, v'}^R(i-1, j-1) \cdot K^R(ch(v, i), ch(v', j))$$

木カーネルの計算困難性：順序なし木には拡張不可能です

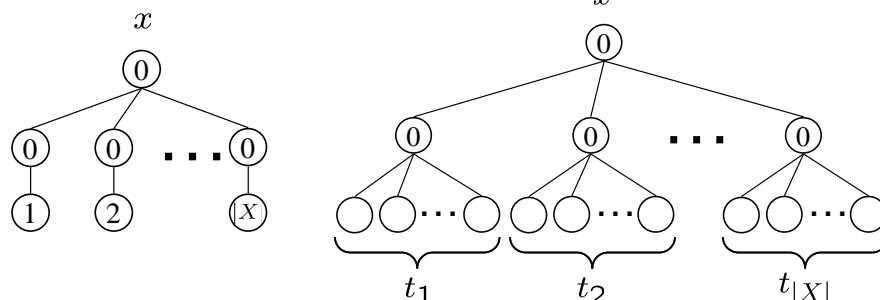
- 順序木カーネルは、Collinsらによる構文解析木カーネル（計算量同じ）の一般化
→ より一層の一般化は可能か？ NO
- 木のクラスとして、順序なしの木にすると、カーネル計算が #P完全になる

- 証明は2つのステップからなる
- 1. #PERFECT MATCHINGS（#P完全性）から、
「大きさ $|V|$ の部分木のみ限定した木カーネル計算」
に帰着することでこの問題の #P完全性を示す
 - ▶ $|V|$ は2つのうち小さいほうの木の大きさ
 - 2つめの木のほうが大きい ($|V|$ より $|V'|$ が大きい) とする
- 2. Cookの還元： 所望の木カーネルを多項式時間で計算するオラクルがあれば、上の問題が多項式時間でとけることを示すことで、矛盾を示す
⇒ 木カーネル計算の #P完全性が示される

木カーネルの計算困難性の証明の概略（1）

1. 「大きさ $|V|$ の部分木のみ限定した部分構造をつかったカーネル」を考えると #P完全
 - ▶ 「大きさ $|V|$ の部分木のみ限定した部分構造をつかったカーネル」は x が x' に埋め込まれる回数を数えるのに等しい
 - 埋め込みのチェックは簡単にできるので #P に入る
 - ▶ 2部グラフ $G = (X \cup Y, E)$ から導かれる図のような2つの木 x, x' を考える
 - 2つの木の定義： X の i 番ノードと Y の j 番ノードの間に枝があるとき、 t_j に i ラベルのノードがある
 - ▶ 「完全マッチングの数を数える」問題から帰着する
 - x が x' に埋め込まれる回数を数えることになる、

#PERFECT MATCHINGS(G)
 入力: 2部グラフ $G=(X \cup Y, E)$, ただし $|X|=|Y|$
 出力: G の完全マッチングの総数



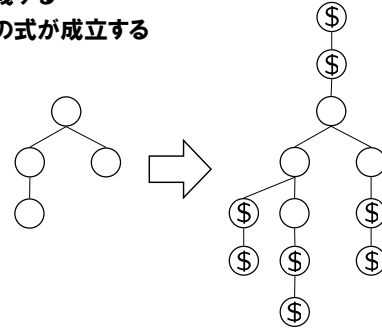
木カーネルの計算困難性の証明の概略 (2)

2. Cookの還元: 所望の木カーネルを計算するオラクルがあれば、
前述の問題が多項式時間でとけることを示すことで、矛盾を示す ([Vadhan01]の技を使用)

- 長さ m のダミーの鎖を全てのノードにつけた木 T_m を定義する
- T_m と T'_m の間の木カーネル $K(T_m, T'_m)$ を考えると、次の式が成立する

$$K(T_m, T'_m) = C_0 + \sum_{n=1}^{|V|} C_n(1 + |V|^n)$$

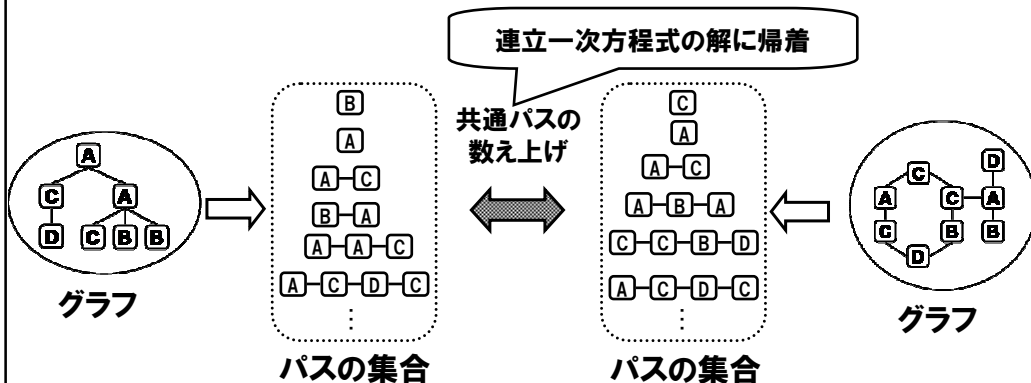
- C_i は部分構造サイズを i に限定したカーネル
 - $C_{|V|}$ が前述の問題の解
- $K(T_m, T'_m)$ が求まるとすると、
 C_i が連立方程式の解としてもとまってしまう → 矛盾



T_m ($m=2$ の場合)

[研究成果 2] グラフに対するカーネル関数の設計を行いました

- グラフカーネルでは、カーネル関数を
 - 部分構造 = パス
 - 数え方 = 2つのグラフからランダムウォークによって共通のパスが生成される個数によって定義する
- 難しいところ: パスは無限個ありうる
 - 解決法: 数え上げの計算を連立一次方程式に帰着することで計算可能になる



H. Kashima, K. Tsuda and A. Inokuchi: Marginalized Kernels Between Labeled Graphs, In Proc. 20th International Conference on Machine Learning (ICML), 2003
Hisashi Kashima, Koji Tsuda and Akihiro Inokuchi: Kernels for Graphs, In Kernel Methods in Computational Biology, MIT Press, 2004

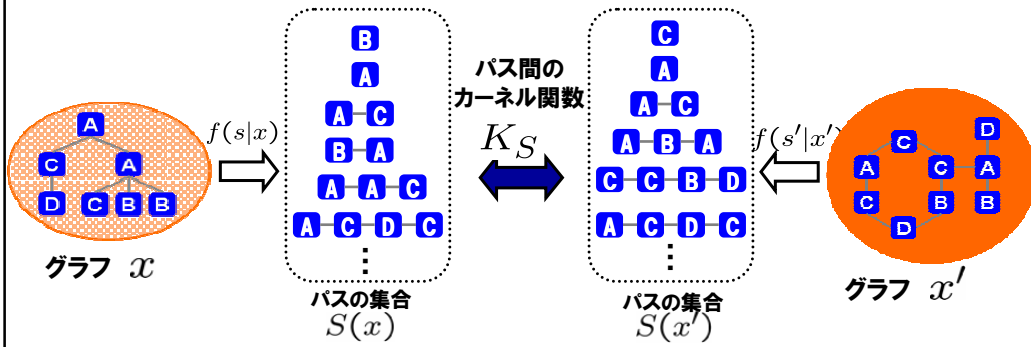
グラフ・カーネルの定義: ランダムウォークによって生成されるパスを用います

- 部分構造の集合 $S(x)$ として何をを使うか?

- 部分グラフ? ⇒ **ダメ** (計算困難)
- グラフ上のランダム・ウォークによって生成されるパス(無限個)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 ($0 < \lambda < 1$)

ホントはランダムウォークの確率だけ簡単のため

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



グラフ・カーネルの計算は、連立方程式に帰着され、効率的に行えます

- ランダム・ウォークによ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

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

- 再帰的表現に持ち込むことで効率的に計算可能になる

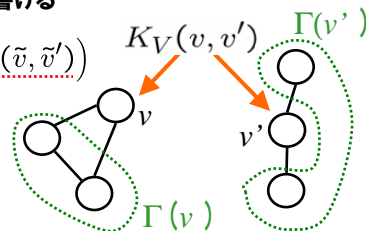
- $S_v(x)$: ノード v で終わるパスの集合 ($S(x) = \cup_{v \in V} S_v(x)$)
- ノード対 (v, v') で終わるパスのみに注目したときのカーネル K_V の和に分解

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

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

- $K_V(v, v')$ は近隣ノード同士の K_V の和によって再帰的に書ける

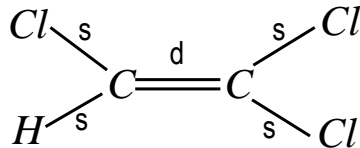
$$K_V(v, v') = \lambda^2 K_\Sigma(v, v') \left(1 + \sum_{\tilde{v} \in \Gamma(v)} \sum_{\tilde{v}' \in \Gamma(v')} \lambda^2 K_V(\tilde{v}, \tilde{v}') \right)$$



- 連立方程式を解けばカーネルが計算できる (多項式時間)

応用：グラフカーネルによる化合物の毒性予測

- 指数時間の他手法（パタンマイニング）に匹敵する性能が得られた



化合物のグラフ表現

パタンマイニング

MinSup	MM	FM	MR	FR
0.5%	60.1%	57.6%	61.3%	66.7%
1%	61.0%	61.0%	62.8%	63.2%
3%	58.3%	55.9%	60.2%	63.2%
5%	60.7%	55.6%	57.3%	63.0%
10%	58.9%	58.7%	57.8%	60.1%
20%	61.0%	55.3%	56.1%	61.3%

$p_\theta(v)$	MM	FM	MR	FR
0.1	62.8%	61.6%	58.4%	66.1%
0.2	63.4%	63.4%	54.9%	64.1%
0.3	63.1%	62.5%	54.1%	63.2%
0.4	62.8%	61.9%	54.4%	65.8%
0.5	64.0%	61.3%	56.1%	64.4%
0.6	64.3%	61.9%	56.1%	63.0%
0.7	64.0%	61.3%	56.7%	62.1%
0.8	62.2%	61.0%	57.0%	62.4%
0.9	62.2%	59.3%	57.0%	62.1%

グラフカーネル

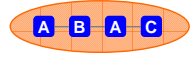
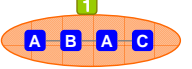
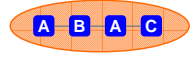
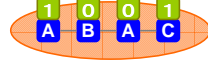
これらの研究はその後、さまざまな発展を遂げています

- 順序木カーネル
 - ▶ 曖昧マッチングのとりこみ [久保山ら, 2006]
 - ▶ 部分構造をパスに限定した高速化 [Kuboyama et al., 2006, 2007]
 - ▶ 糖鎖構造分類への応用 [Kuboyama et al., 2006]
- グラフカーネル
 - ▶ ランダムウォークの設計による高精度化 [Mahe et al., 2004]
 - ▶ 疎行列を利用した各種高速化 [Vishwanathan et al., 2006]
 - ▶ Cyclic Pattern [Horvath et al., 2004]、Shortest Path [Borgwardt & Kriegel., 2005] などの近似による高速化
 - ▶ タンパク質立体構造分類への適用 [Borgwardt et al., 2005]
- より複雑な問題への適用
 - ▶ 構造データのラベル付け問題への適用 [Kashima et al., 2004]

その後の発展の一例...

[研究成果 3] 構造データのラベル付け問題への適用 を行いました

- ラベル付け問題は、構造データの各要素に対してラベルを割り当てる予測問題
- ラベル付け問題の例：
 - ▶ 形態素解析、固有表現抽出、タンパク質の2次構造予測、遺伝子領域の予測、...
- 各要素の分類問題と考えることでカーネル法を適用できる

	入力	出力
分類問題	 構造データ	 構造データのクラス
ラベル付け問題	 構造データ	 各要素のラベル

H. Kashima and Y. Tsuboi: Kernel-Based Discriminative Learning Algorithms for Labeling Sequences, Trees and Graphs, In Proc. 21st International Conference on Machine Learning (ICML), 2004.

応用：自然言語文からの情報抽出

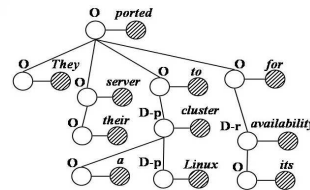
- 局所的な情報 (bi-gramモデル) に基づく方法を上回る性能が得られた
 - ▶ 固有表現抽出
 - 人名、組織名、場所名などを示すフレーズの抽出
 - ▶ 製品使用情報抽出
 - ある製品を使っているという事実を抽出
 - {製品名/企業名/数量/理由} × {導入/導入検討中/導入しない} など
 - 構文解析木構造も用いることで精度を向上できる

表 1 固有表現抽出結果 (括弧内は標準偏差)

	精度	適合率	再現率	F1
配列カーネル	88.7% (3.4)	49.0% (6.0)	23.1% (8.1)	30.5 (6.7)
HM パーセプトロン	80.2% (11.5)	23.8% (14.6)	17.9% (3.0)	18.6 (5.2)

表 2 製品使用情報抽出結果 (括弧内は標準偏差)

	精度	適合率	再現率	F1
SEQUENCE KERNEL	89.7% (2.0)	52.2% (9.5)	29.6% (4.0)	37.5(4.1)
TREE KERNEL	89.9% (2.7)	51.4% (10.9)	32.5% (14.4)	38.9 (12.1)
HM-PERCEPTRON	89.7%(1.8)	51.5%(8.5)	24.0%(21.4)	28.9(20.3)



前半のまとめ：内部構造の扱いに関し、カーネル法の研究を行いました

■ 構造データのカーネル関数を、表現力と計算の速さのバランスをとりながら、うまく設計しました

- ▶ [研究成果 1] ラベル付順序木カーネル
 - 順序木に対するカーネルを、部分木を用いて定義しました
 - 動的計画法を用いた効率的なアルゴリズムを設計しました
 - 順序なし木に対するカーネル関数の計算困難性を示しました
- ▶ [研究成果 2] グラフ・カーネル
 - グラフに対するカーネルを、ランダムウォークによって生成されるパスを用いて定義しました
 - 連立一次方程式に帰着することで効率的なアルゴリズムを設計しました
- ▶ [研究成果 3] ラベル付けカーネルマシン
 - ラベル付け問題に、畳み込みカーネル法を適用しました

外部構造の扱いについての研究成果

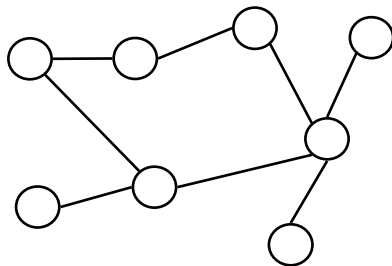
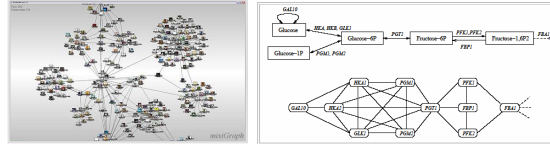
外部構造の解析としては、リンク予測の問題に取り組みました

▪ **リンク予測問題**: 部分的に観測されているネットワーク構造から、残りの構造を推定する問題

▶ 「リンクマイニング」の主要タスクのひとつ

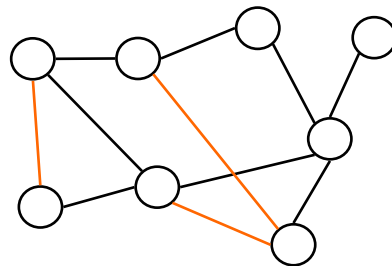
▶ 例:

- 生体ネットワークの構造予測
- SNSにおける友人の推薦
- テロリストの「隠れた繋がり」の発見



部分的に観測される構造

補完
⇒



予測される構造

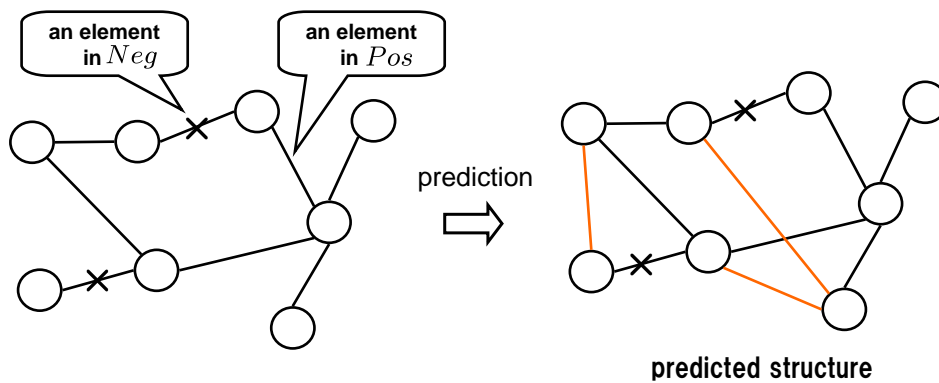
リンク予測問題は、グラフ構造の補完問題として定義できます

▪ **入力**: 一部が欠けたネットワーク構造

- ▶ リンクありのノードペア: Pos
- ▶ リンクのないノードペア: Neg

▪ **出力**:

- ▶ リンクの有無が未知のノードペアについてのリンク予測



リンク予測の応用

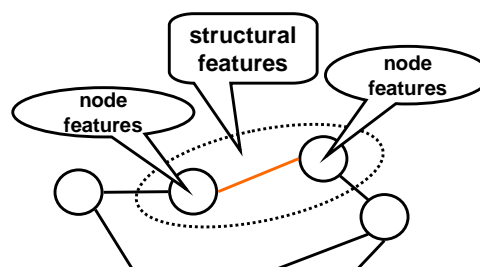
- 応用:
 - ▶ 生体ネットワークの構造予測
 - ▶ ソーシャルネットワークでの推薦や関係の予測
- ネットワーク構造をもったデータの例:

ネットワーク	ノード	リンク
WWW	Webページ	ハイパーリンク
社会ネットワーク	人 コミュニティ	友人関係 所属
生体ネットワーク	遺伝子 タンパク質	制御 相互作用

リンク予測に用いることのできる情報にはノード情報と構造情報があります

- リンク予測には、ノード情報と構造情報の2つの情報を用いることができます
- ノード情報: ノード自身のもつ情報
 - ▶ ソーシャルネットワークでは、各人(=ノード)は、住所や年齢などの個人情報をもつ
 - ▶ タンパク質ネットワークでは、各タンパク質(=ノード)は、配列情報や発現情報などをもつ
- 構造情報: (ノードペアの周辺の)構造のもつ情報
 - ▶ ソーシャルネットワークで、友達の友達は友達である可能性が高い

今回は、構造情報のみに注目



[研究成果 4] ネットワーク構造の時間変化モデルの学習によって、ネットワークの構造情報のみからリンク予測を行う手法を提案しました

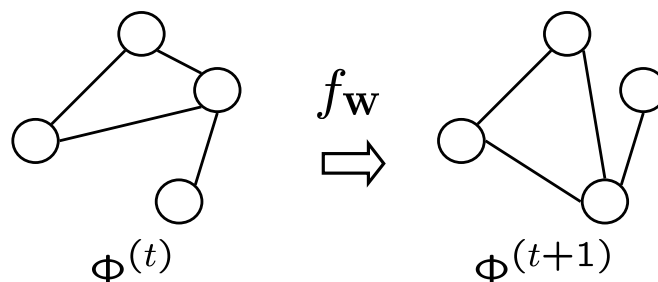
- ネットワーク構造の時間変化モデルを考え、これをもとにリンク予測を行う
 - ▶ ノードからノードへの枝の「コピー & ペースト」が確率的におこるモデルを仮定する
- 観測されるネットワークは、この時間変化モデルの定常分布から生成されたと仮定して、パラメータを学習、残りの部分を予測する
 - ▶ この問題を近似的に解く学習アルゴリズムを提案する



H. Kashima and N. Abe: A Parameterized Probabilistic Model of Network Evolution for Supervised Link Prediction, *IEEE International Conference on Data Mining (ICDM)*, 2006.
 鹿島 久嗣, 安倍直樹: ネットワーク構造の確率的な時変モデルに基づく教師ありリンク予測, *人工知能学会論文誌*, Vol.22, No.2, 2007.

まず、リンク予測の土台として、ネットワークの構造が時間的に変化していくようなモデルを考えます

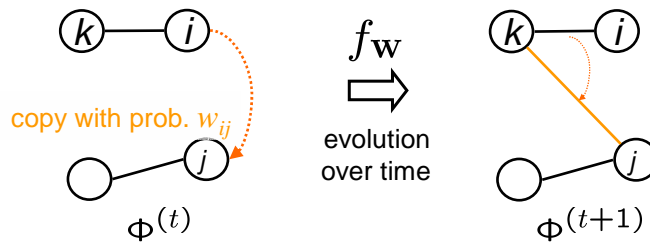
- リンク予測を考えるにあたり、まず、リンクがどのように生成されるかというモデルを考える
- $\Phi(t)$: 時刻 t におけるネットワーク構造
 - ▶ リンクラベル: $\phi^{(t)}(i, j) = \begin{cases} 1 & \text{if a link exists between } i \& j \\ 0 & \text{otherwise} \end{cases}$
- f_w : ネットワークの構造変化モデル (パラメータ w をもつ)
- $\Phi(t)$ と f_w がわかると、次の時刻 $t+1$ でのネットワーク構造 $\Phi^{(t+1)}$ がきまる



具体的なモデルとして、
「コピー＆ペースト」に基づくネットワーク構造の進化モデルを仮定します

- 各時点で、ネットワークのどこかで、あるノード i からあるノード j へリンクラベルがコピーされる
 - ノードペア (i, j) を確率 w_{ij} で選択 (パラメータには確率的な制約 $\sum_{i,j} w_{i,j} = 1, w_{i,j} \geq 0$ がある)
 - 別のノード k を一様にランダムに選ぶ (j 以外)
 - リンクラベル $\phi^{(t)}(i, k)$ がノードペア (j, k) にコピーされる

$$\phi^{(t+1)}(j, k) := \phi^{(t)}(i, k) \quad \phi^{(t)}(i, j) = \begin{cases} 1 \\ 0 \end{cases}$$

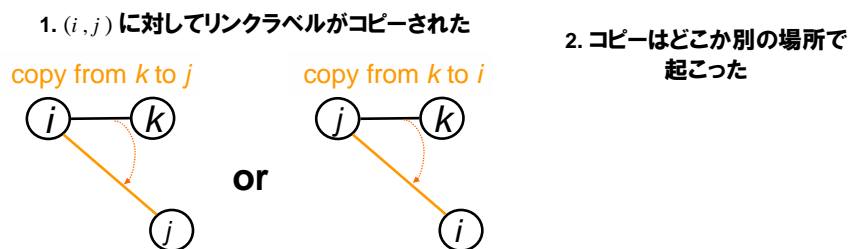


- 類似のモデルがKleinbergらによってWWWの構造進化モデルとして提案されている
 - また、生体ネットワークにも同じようなモデルが当てはまると言われている

リンクごとにみたときの、リンクの有無の時間変化は形式的に書けます

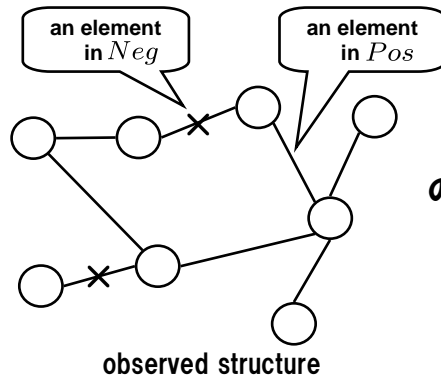
- 時刻 $t+1$ において i と j の間にリンクが存在する条件付(周辺)確率 $\Pr[\phi^{(t+1)}(i, j) = 1 | \Phi^{(t)}]$ は、次の2つの可能性から決まる
 - 時刻 t において、 (i, j) に対してどこからリンクラベルがコピーされた
 - 時刻 t において、コピーはどこか別の場所で起こった (つまり (i, j) に対しては何も起こらなかった)

$$\Pr[\phi^{(t+1)}(i, j) = 1 | \Phi^{(t)}] = \underbrace{\frac{1}{|V|-1} \left(\sum_{k \neq i, j} w_{kj} \phi^{(t)}(k, i) + w_{ki} \phi^{(t)}(k, j) \right)}_{1. (i, j) \text{ に対してリンクラベルがコピーされた}} + \underbrace{\left(1 - \frac{1}{|V|-1} \sum_{k \neq i, j} w_{kj} + w_{ki} \right) \phi^{(t)}(i, j)}_{2. \text{コピーはどこか別の場所で起こった}}$$



技術的な課題: 履歴なし、部分構造のみからの構造推定を解決する必要があります

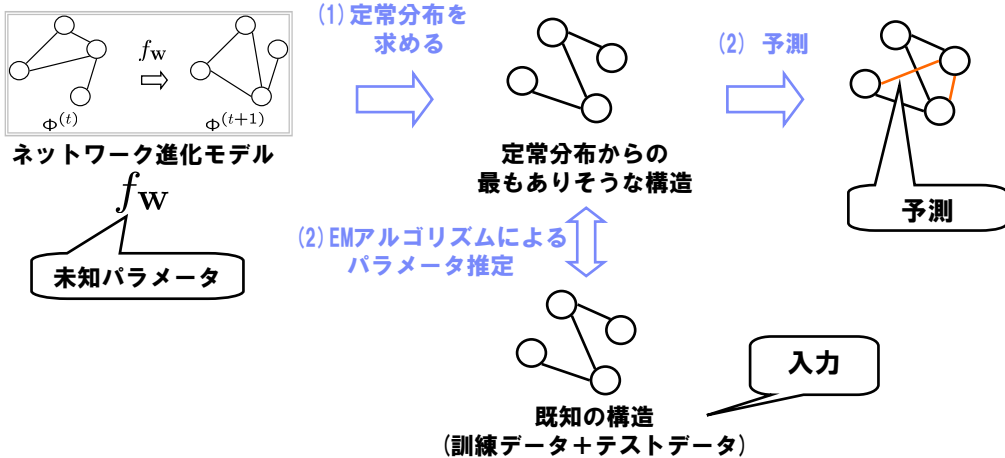
- ネットワーク構造変化の履歴が与えられていない
- わかっているのは、現在のネットワークの部分構造のみ



のみから推論する必要がある

戦略: ネットワーク構造進化モデルの定常分布を考え、既知のネットワーク構造にフィットさせます

- 2つのステップ:
 1. ネットワーク進化モデルからネットワーク構造の定常分布を導く
 2. 定常分布を、ネットワークの既知の部分にフィットさせる



各ステップの詳細:

(1) ネットワーク構造進化モデルから定常分布を求める

▪ 現在のネットワークが、定常分布の最もありそうな構造と一致すると仮定

▪ 「コピー-&ペーストモデル」で、 $t \rightarrow \infty$ とする

$$\Pr[\phi^{(t+1)}(i, j) = 1 | \phi^{(t)}] = \frac{1}{|V|-1} \left(\sum_{k \neq i, j} w_{kj} \phi^{(t)}(k, i) + w_{ki} \phi^{(t)}(k, j) \right) + \left(1 - \frac{1}{|V|-1} \sum_{k \neq i, j} w_{kj} + w_{ki} \right) \phi^{(t)}(i, j)$$

↓ $t \rightarrow \infty$

▪ (周辺分布の)定常分布は...

▶ $\rho(i, j) := \Pr[\phi^{(\infty)}(i, j) = 1]$ とおくと

$$\rho(i, j) = \frac{\sum_{k \neq i, j} w_{kj} \rho(k, i) + w_{ki} \rho(k, j)}{\sum_{k \neq i, j} w_{kj} + w_{ki}}$$

▶ この連立方程式を満たす $\rho(i, j)$ が、ネットワーク進化モデルを代表する「平均的な」構造であると仮定する

各ステップの詳細:

(2) 定常分布をネットワークの既知の部分にフィットさせる

▪ (周辺分布の) 定常分布は

$$\rho(i, j) = \frac{\sum_{k \neq i, j} w_{kj} \rho(k, i) + w_{ki} \rho(k, j)}{\sum_{k \neq i, j} w_{kj} + w_{ki}} \quad \text{ただし、} \quad \rho(i, j) := \Pr[\phi^{(\infty)}(i, j) = 1]$$

↓ 目的関数に代入

▪ 目的関数: リンクラベルごとの対数尤度の和を最大化

リンク既知のノードペア $\sum_{(i, j) \in Pos \cup Neg} \phi(i, j) \log \rho(i, j) + (1 - \phi(i, j)) \log(1 - \rho(i, j))$

リンク未知のノードペア $\sum_{(i, j) \notin Pos \cup Neg} \rho(i, j) \log \rho(i, j) + (1 - \rho(i, j)) \log(1 - \rho(i, j))$

▪ リンク未知のノードペア $(i, j) \notin Pos \cup Neg$ にたいする $\rho(i, j) := \Pr[\phi^{(\infty)}(i, j) = 1]$ が最終的な予測

最適化問題の解法: パラメータ推定と未知変数推定を交互に行います

リンク既知のノードペア $\sum_{(i,j) \in Pos \cup Neg} \phi(i,j) \log \rho(i,j) + (1 - \phi(i,j)) \log(1 - \rho(i,j))$

リンク未知のノードペア $\sum_{(i,j) \notin Pos \cup Neg} \rho(i,j) \log \rho(i,j) + (1 - \rho(i,j)) \log(1 - \rho(i,j))$

- と を、交互に最適化する

- を固定して、を求める

- 現在のパラメータに基づいて、連立方程式をとく

$$\rho(i,j) = \frac{\sum_{k \neq i,j} w_{kj} \rho(k,i) + w_{ki} \rho(k,j)}{\sum_{k \neq i,j} w_{kj} + w_{ki}}$$

- を固定して、を最大化する

- 勾配の計算を行うのに、連立方程式をたくさん解く必要がある → 思い切って近似

$$\frac{\partial \rho(i,j)}{\partial w_{lm}} = (\rho(k,i) \text{ や } \rho(k,j) \text{ にまつわる項}) + (\frac{\partial \rho(k,i)}{\partial w_{lm}} \text{ や } \frac{\partial \rho(k,j)}{\partial w_{lm}} \text{ にまつわる項})$$

- 実装上は、データをひとつずつ処理する逐次型の学習を行う

提案手法はEMアルゴリズムの近似解法としても解釈できます

- 目的関数は、訓練データの同時分布の対数尤度 $\log \Pr(\Phi^L | W)$ とする

- Φ^L : 訓練データ、 Φ^U : テストデータ とする

- EMアルゴリズム: 対数尤度の下界を最大化するようにパラメータ更新 ($\widehat{W} \rightarrow W$)

- 提案手法では、この下界を平均場近似したものを最大化している

$$Q(W, \widehat{W}) := \sum_{\Phi^U} P(\Phi^U | \Phi^{(\infty)L}, \widehat{W}) \log P(\Phi^L, \Phi^U | W)$$

$$\approx \sum_{\Phi^U} \prod_{(i,j) \in E^U} P(\phi(i,j) | \Phi^L, \widehat{W}) \log \prod_{(k,l) \in E} P(\phi(k,l) | W)$$

$$= \sum_{(i,j) \in E^U} (\rho(i,j | \Phi^L, \widehat{W}) \log \rho(i,j | W) + (1 - \rho(i,j | \Phi^L, \widehat{W})) \log(1 - \rho(i,j | W)))$$

$$+ \sum_{(i,j) \in E^L} \phi(i,j) \log \rho(i,j | W) + (1 - \phi(i,j)) \log(1 - \rho(i,j | W))$$

平均場近似

提案手法の目的関数

- 最適性は保証されない

- 下界自体が近似

- 近似された下界の最適化においても、微分を近似している

実験：既知のリンク指標との比較

- リンク指標 := 構造情報に基づく、リンクの確からしさ
 - ▶ 社会ネットワーク分析において、いろいろ提案されている
 - 一部、情報検索の分野でも
 - ▶ 例：Common neighbors指標 は、2つのノードの共通の隣接ノード数によって定義される
 - 「友達の友達は、たぶん友達」
- リンク指標の大きい順にリンクを予測する

比較対象に用いたリンク指標の例

- Common neighbors: 共通の隣接ノード数

$$\text{common neighbors} := |\Gamma(i) \cap \Gamma(j)|$$

$\Gamma(i)$ はノード i の隣接ノード集合

- 重み付き common neighbors:

$$\text{Adamic/Adar} := \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log |\Gamma(k)|}$$

$$\text{Jaccard's coefficient} := \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|}$$

(used in information retrieval)

- 長距離 common neighbors:

$$\text{Katz}_\beta := \sum_{l=1}^{\infty} \beta^l |\text{paths}_{i,j}^{(l)}|$$

$\text{paths}_{i,j}^{(l)}$ はノード i からノード j への長さ l のパスの数

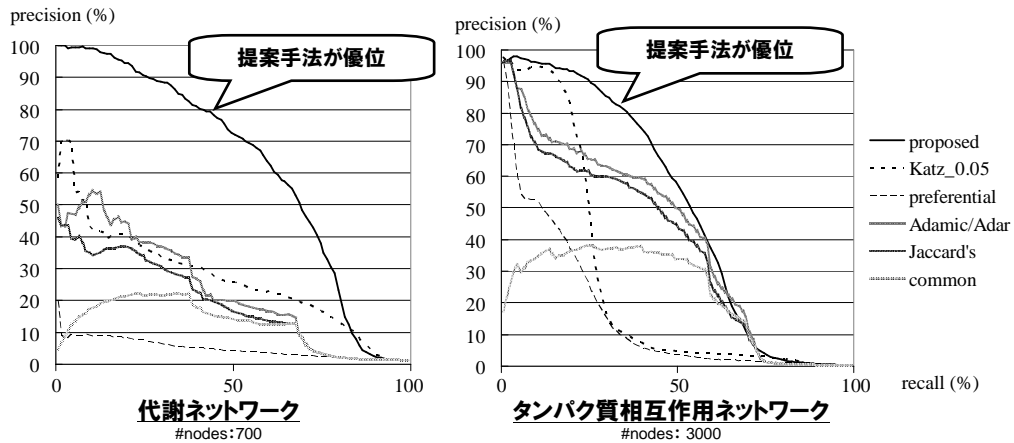
- 優先的選択

- ▶ 隣接ノード数の積

$$\text{preferential attachment} := |\Gamma(i)| \cdot |\Gamma(j)|$$

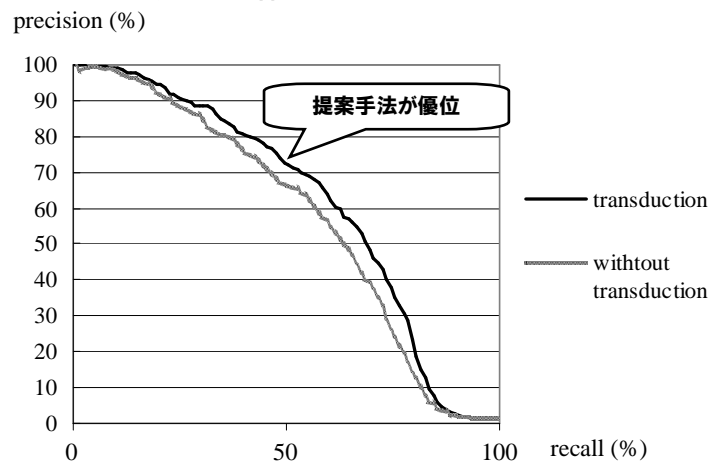
実験結果：生体ネットワークにおいては良好な予測性能でした

- 代謝ネットワークとタンパク質相互作用ネットワークを用いた構造予測実験
- 提案手法が、その他のリンク指標を上回る性能
 - ▶ 3-fold cross validation
- ただし、文献ネットワークでは性能悪し
 - ▶ 提案したモデルが合っているかどうかによると思われる



近似によって無理に解いたメリットはあったのか？ Yes

- 数々の計算の煩雑さの原因は、未知変数を学習のなかで用いていたことによる
 - 頑張った甲斐はあったのか？
- 未知変数を用いない場合（平均場近似以外は厳密に解ける）と比較して、精度が向上します
 - ▶ 観測されないリンクラベルは用いずに学習する



後半のまとめ：外部構造の扱いについて、
確率的なリンク予測手法の研究を行いました

- リンク予測問題への確率的アプローチ
- [研究成果 4] ネットワークの進化モデルに基づくリンク予測
 - ▶ ネットワーク構造が、コピーによって構造を変えていくようなモデルを提案しました
 - ▶ 「構造進化モデルの定常分布を構造の既知の部分にフィットさせる」という考え方によって、リンク予測問題を解くアプローチを提案しました
 - ▶ 近似的なEMアルゴリズムによる推定アルゴリズムを設計しました
- 課題：提案アプローチの一般性と限界
 - ▶ どのようなデータに対して本手法が有効か？
 - ▶ 近似した部分を取り除くことで精度向上が望めるか？

本発表のまとめ：構造データを扱う機械学習手法の研究

- 構造をもったデータを効率的に解析するための手法を研究しました
- 特にカーネル法とよばれる手法を用いた、内部構造をもつデータの解析手法に大きく貢献しました
- また、外部構造をもつデータの解析については、リンク予測の研究を行い研究成果を得ました

- 今後も、ネットワーク構造予測の研究を推し進め、
 - ▶ スケールフリーネットワークモデルとカーネル法の融合
 - ▶ 実際の生体ネットワーク構造予測における知見の発見を目指します

本論文の内容に含まれる発表文献

- ジャーナル論文
 - ▶ 鹿島 久嗣, 安倍直樹:
ネットワーク構造の確率的な時変モデルに基づく教師ありリンク予測,
人工知能学会論文誌, Vol.22, No.2, 2007.
 - ▶ 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.
- 査読つき国際会議論文
 - ▶ Hisashi Kashima and Naoki Abe:
A Parameterized Probabilistic Model of Network Evolution for Supervised Link Prediction,
In Proc. *IEEE International Conference on Data Mining (ICDM)*, 2006.
 - ▶ Hisashi Kashima and Yuta Tsuboi:
Kernel-Based Discriminative Learning Algorithms for Labeling Sequences, Trees and Graphs,
In Proc. *21st International Conference on Machine Learning (ICML)*, 2004.
 - ▶ Hisashi Kashima, Koji Tsuda and Akihiro Inokuchi:
Marginalized Kernels Between Labeled Graphs,
In Proc. *20th International Conference on Machine Learning (ICML)*, 2003.
 - ▶ Hisashi Kashima and Teruo Koyanagi:
Kernels for Semi-Structured Data,
In Proc. *19th International Conference on Machine Learning (ICML)*, 2002.
- 著書(章)
 - ▶ Hisashi Kashima, Koji Tsuda and Akihiro Inokuchi: Kernels for Graphs,
In *Kernel Methods in Computational Biology*, MIT Press, 2004.

本論文の内容に関連する発表文献

- ジャーナル論文
 - ▶ Tetsuji Kuboyama, Hisashi Kashima, Kiyoko F. Aoki-Kinoshita, Kouichi Hirata, Hiroshi Yasuda:
A Spectrum Tree Kernel,
人工知能学会論文誌, Vol.22, No.2, 2007.
- 査読つき国際会議論文
 - ▶ Tetsuji Kuboyama, Kouichi Hirata, Kiyoko F. Aoki-Kinoshita, Hisashi Kashima and Hiroshi Yasuda:
A Gram Distribution Kernel Applied to Glycan Classification and Motif Extraction,
In Proc. *17th International Conference on Genome Informatics (GIW2006)*, 2006.
 - ▶ Tetsuji Kuboyama, Hisashi Kashima, Kiyoko F. Aoki-Kinoshita, Koichi Hirata and Hiroshi Yasuda:
A Spectrum Tree Kernel,
In Proc. *the International Workshop on Data-Mining and Statistical Science (DMSS2006)*, 2006.
 - ▶ Tetsuji Kuboyama, Kilho Shin and Hisashi Kashima:
Flexible Tree Kernels Based on Counting the Number of Tree Mappings,
In Proc. *Workshop on Mining and Learning* (held with ECML/PKDD 2006), 2006
 - ▶ Akihiro Inokuchi and Hisashi Kashima:
Mining Significant Pairs of Patterns from Graph Structures with Class Labels,
In Proc. *3rd IEEE International Conference on Data Mining (ICDM2003)*, 2003.

ありがとうございました

