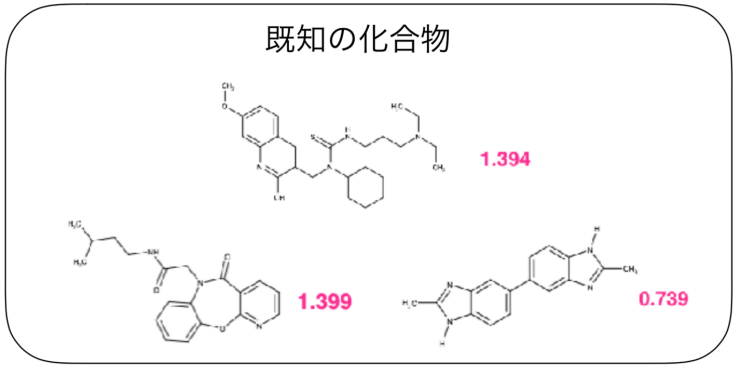
北海道大学

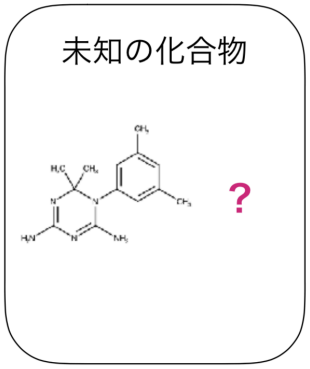
入力表現の適応的選択を伴うグラフ畳み込みネットワーク学習

情報処理学会第81回全国大会
2019年3月15日

北海道大学大学院情報科学研究院
◯菊地 翔馬　瀧川一学

本研究の背景・目的



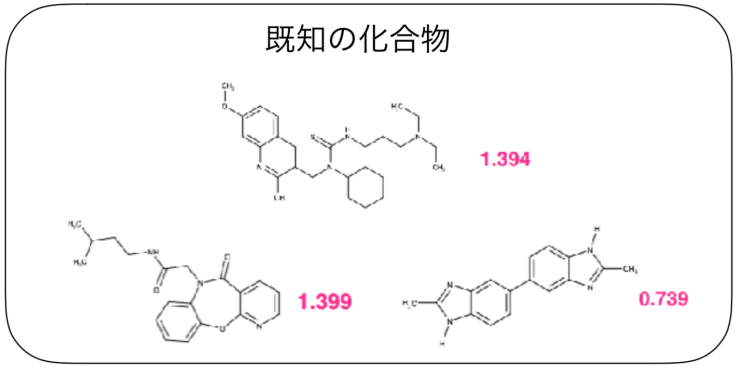


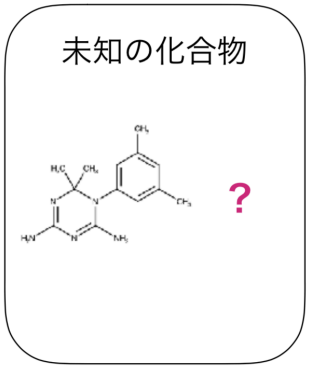
未知の化合物の溶解度を予測したいタスク

・有機低分子の性質の予測モデリングの研究がされている

・医薬品開発など、実験コストの削減ができる

本研究の背景・目的





・有機低分子の性質の予測モデリングの研究がされている

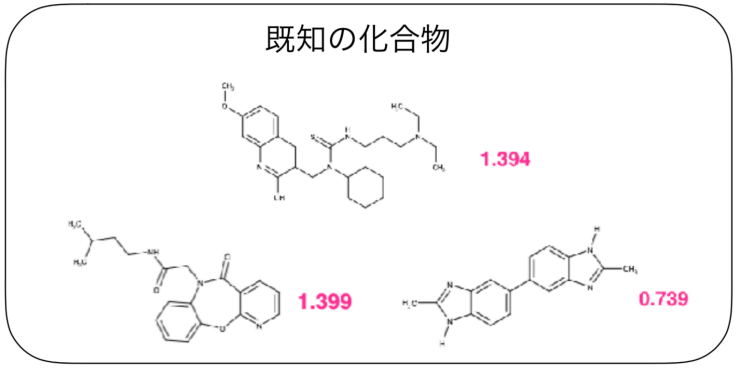
・医薬品開発など、実験コストの削減ができる

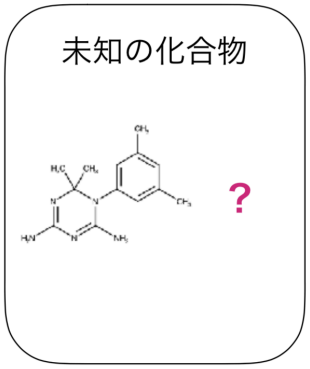
未知の化合物の溶解度を予測したいタスク

情報科学：グラフ構造に着目し、教師つき学習で予測

化学：化学構造や化学的知識を用いて、化合物の性質を予測

本研究の背景・目的





・有機低分子の性質の予測モデリングの研究がされている

・医薬品開発など、実験コストの削減ができる

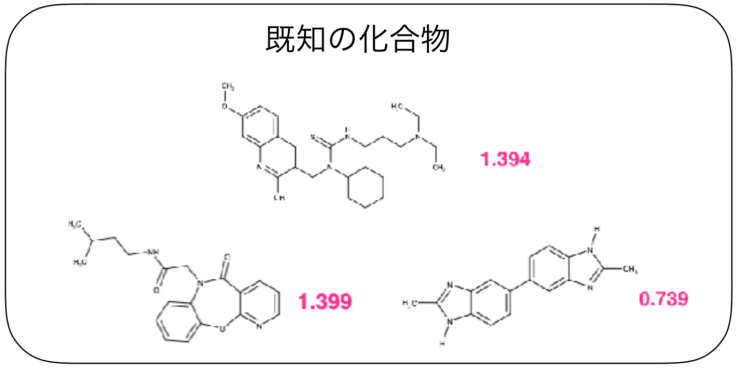
未知の化合物の溶解度を予測したいタスク

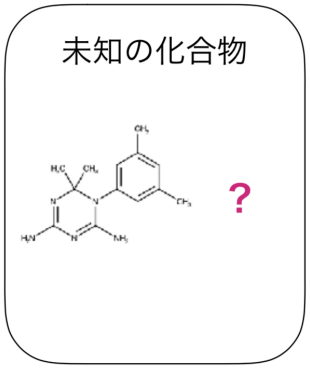
我々はこっち

情報科学：グラフ構造に着目し、教師つき学習で予測

化学：化学構造や化学的知識を用いて、化合物の性質を予測

本研究の背景・目的





・有機低分子の性質の予測モデリングの研究がされている

・医薬品開発など、実験コストの削減ができる

未知の化合物の溶解度を予測したいタスク

我々はこっち

情報科学：グラフ構造に着目し、教師つき学習で予測

化学：化学構造や化学的知識を用いて、化合物の性質を予測

グラフ構造と**化学的知識**の関係性の理解が難しい

適応的に予測に必要な知識を選択する学習モデルを提案

発表の流れ

1.前提知識

- グラフ回帰問題
- 分子グラフとラベルづけ
- 既存研究
- 畳み込みネットワーク

2.提案手法

3.実験

4.まとめ

グラフ回帰(分類)問題

グラフ表現 G とそれに対応する関連値 y を持つデータセットを用いた教師付き学習

$$\hat{y} = f_\theta(G)$$

1.23

$\hat{y}$

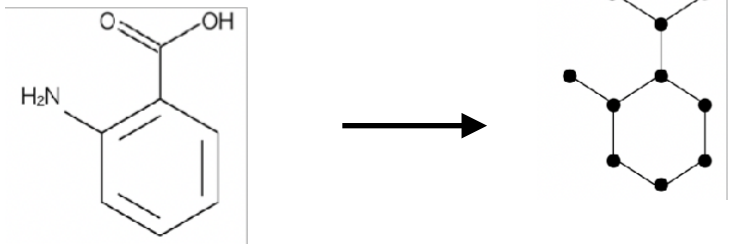
グラフ表現 G

- グラフを数値ベクトルに変換し、回帰（分類）問題として解く
- モデルが持つパラメータθを学習し、入力に対する予測値を計算する
- グラフ表現は頂点や辺に多次元ラベルを持っている

分子グラフ：入力分子のグラフ表現

グラフトポロジー

例）アントラニル酸

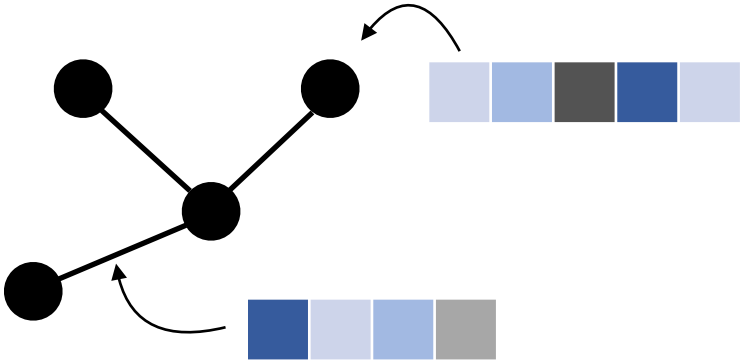


標準的なグラフ構造　　分子グラフ

・原子をノード、結合をエッジとして表現したグラフ

・水素を無視したり、官能基は一つのノードとして見ることもある

頂点と辺の入力表現(多次元ラベル)



・原子や結合の性質を表す「**入力表現**」を多次元ラベルとして、ノードとエッジに持たせておく

既存研究 (*Convolution Networks on Graphs for Learning Molecular Fingerprints)

Neural FingerPrints(NFP) [Duvenaud+, 2015]

グラフ表現 G

→ NFP →

1.23

$\hat{y}$

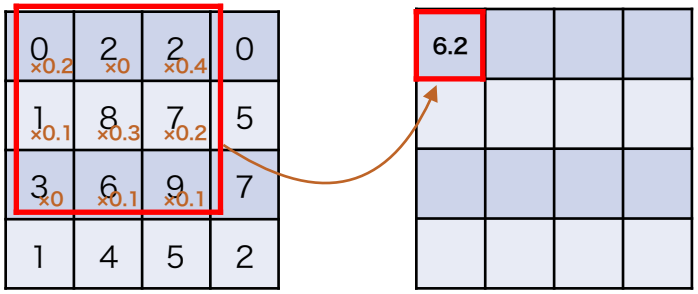
- ニューラルネットワークでグラフの変換から予測まで全て学習
- 畳み込みネットワーク(CNN)をグラフ入力に拡張
- 分子グラフの頂点と辺の**入力表現(多次元ラベル)を事前に決定する必要がある**

畳み込みネットワーク(CNN)

・画像処理などで用いられるCNNをグラフ構造データに適用する

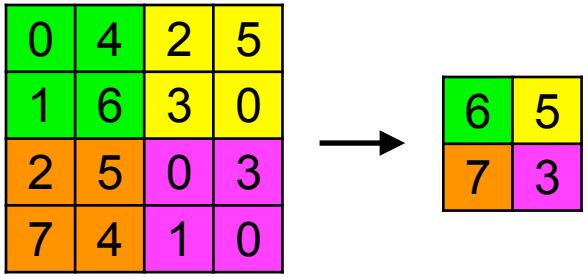
画像(配列)のCNN

畳み込み操作



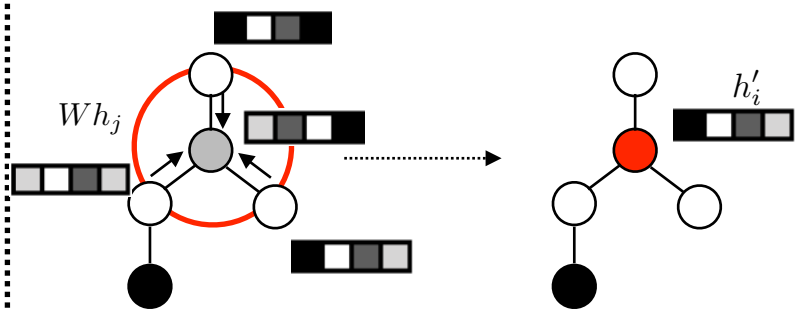
フィルタリングによる変換

プーリング操作

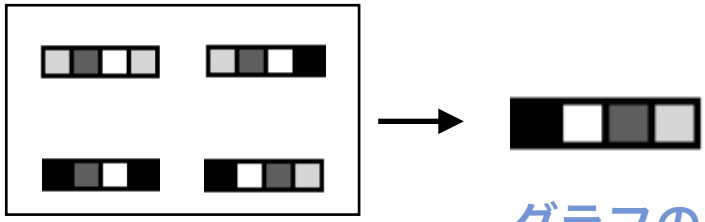


例）Max Pooling

グラフのCNN



隣接するノード情報を用いて変換



集約された
各ノード情報

グラフの
特徴ベクトル

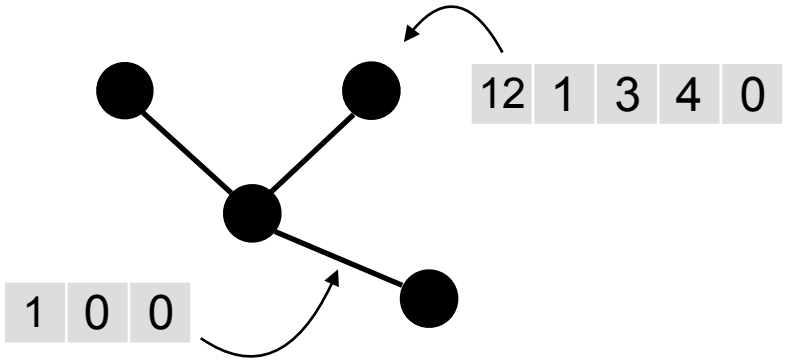
既存研究で用いられた入力表現

・既存研究で用いられた**頂点と辺の入力表現(多次元ラベル)**

◆ノード：原子の**物理**的性質

◆エッジ：結合の種類

エッジラベル
n重結合
共役系
環状



※実際は one-hot encoding

ノードラベル
原子番号
次数
総水素数
価電子
ベンゼン環

入力表現の恣意性

・**しかし**、他にも代表的な情報がある・・・

原子の**物理**的性質

ノードラベル
原子番号
次数
総水素数
価電子
ベンゼン環

原子の**化学**的性質

ノードラベル
ドナー
アクセプター
芳香族
ハロゲン
酸性
塩基性

・データやタスクによっては、入力表現に何を用いるかで、精度に影響するかも

入力表現の恣意性

・**しかし**、他にも代表的な情報がある・・・

原子の**物理**的性質

ノードラベル
原子番号
次数
総水素数
価電子
ベンゼン環

原子の**化学**的性質

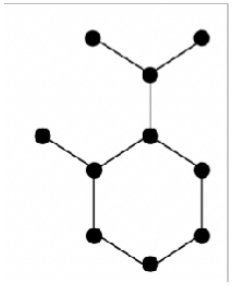
ノードラベル
ドナー
アクセプター
芳香族
ハロゲン
酸性
塩基性

・データやタスクによっては、入力表現に何を用いるかで、精度に影響するかも

・選択するには専門知識が必要・・・

提案モデル

- 適応的に必要な入力表現を重み付き選択する層を導入
- 専門知識がなくても必要な表現を選んでくれる



頂点と辺の
入力表現1

頂点と辺の
入力表現2

・
・
・

頂点と辺の
入力表現K

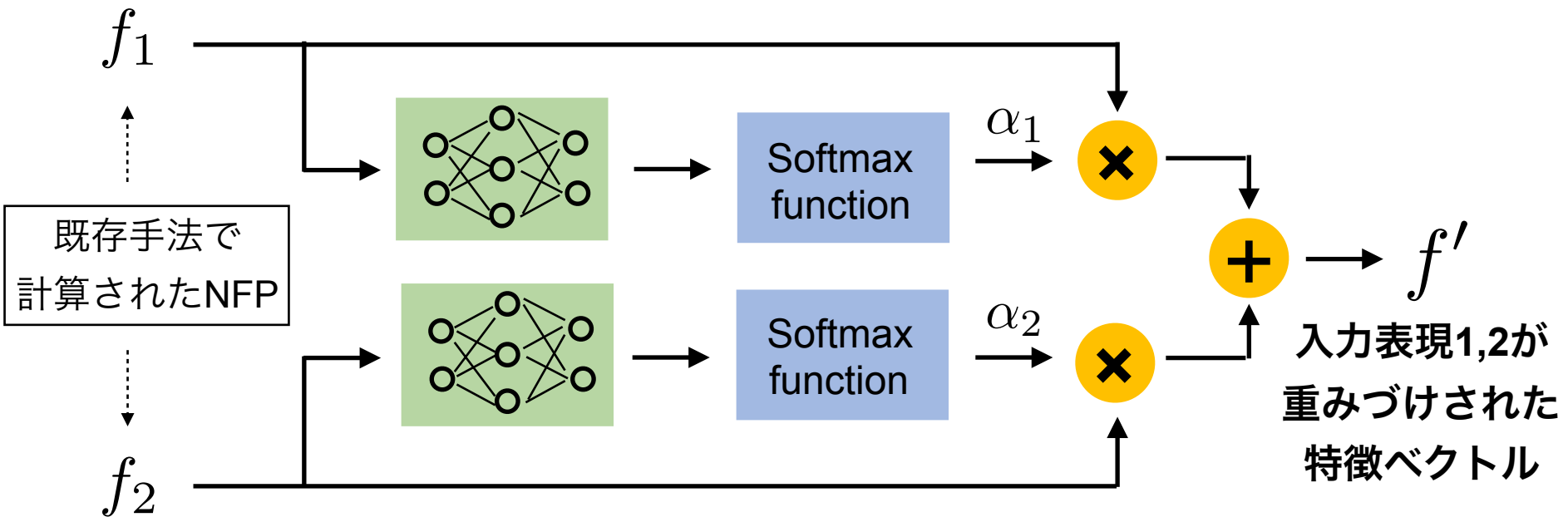
Soft
Attention
Layer

f

各入力表現の重みを
考慮された特徴ベクトル

Soft Attention層の役割

- 着目すべき重みを表す、注目度係数ベクトル α を計算する
- この重みをデータごとに計算することで適応的選択をする



計算機実験

- 実験内容
  - ◆既存手法(NFP)と、提案手法を実装し、提案部分以外は同様の条件で学習し、予測精度が向上することを確認する
- 評価方法
  - ◆既存研究[Duvenaud+, 2015]で用いられていた同様のデータセットを用いて、学習した結果のテストデータに対するRMSEを比較

計算機実験

・分子グラフの辺の入力表現

エッジラベル	サイズ
結合の種類	4
共役系	1
環状	1
合計	6

◆エッジラベルは１種類のみ

計算機実験

・分子グラフの頂点の入力表現

入力表現 1：物理的性質

ノードラベル	サイズ
原子番号	49
次数	6
総水素数	5
価電子	6
ベンゼン環	1
合計	67

入力表現 2：化学的性質

ノードラベル	サイズ
ドナー	1
アクセプター	1
芳香族	1
ハロゲン	1
酸性	1
塩基性	1
合計	6

◆2 種類の異なる表現であるノード情報

◆Soft Attentionを用いて適応的に選択する

計算機実験

- データセット
  - ◆既存研究[Duvenaud, 2015]内で、用いられていたデータと同様
    - 溶解性: 化合物の溶解度 [log Mol/L]のデータセット
    - 薬効：熱帯熱マラリア原虫 P.falciparum の硫化物耐性に対する,試験管内での半数効果濃度 EC_{50}[nM] のデータセット
    - 有機光起電力効果: 有機分子の光起電力効率 [%] の密度汎関数法による計算値データセット

	溶解性	薬効	有機光起電力効果
size	1,144	10,000	29,978
train	700	7,000	20,000
validation	200	1,900	6,000
test	100	1,000	3,000

計算機実験

- 実験設定
  - ◆Soft Attention 層のニューラルネットワークのノード数は(50, 100, 50, 50)
  - ◆epoch数は1000
  - ◆その他、NFPに関わるハイパーパラメータや最適化関数は既存研究と同様の条件
  - ◆既存手法に2つの入力表現を結合し情報量を同等にしたものとも比較

実験結果

- **提案手法による精度の改善を確認**

	溶解性	薬効	有機光起電力効果
平均	4.29 ± 0.40	1.47 ± 0.07	6.40 ± 0.09
NFP+入力表現1	1.09 ± 0.04	1.10 ± 0.03	1.89 ± 0.00
NFP+入力表現2	1.26 ± 0.05	1.12 ± 0.02	2.89 ± 0.02
NFP+入力表現1,2	1.14 ± 0.04	**1.09 ± 0.02**	1.87 ± 0.04
提案法+入力表現1,2	**1.00 ± 0.13**	**1.09 ± 0.00**	**1.68 ± 0.02**

◆情報量が同等でも精度の改善が確認できた

◆適応的に入力表現の選択ができていると考えられる

まとめ

- 適切な入力表現を選択する機構を提案
  - ◆注目度係数を計算するSoft Attention層を導入
  - ◆専門知識がなくても複数の入力表現を適応的に選択可能

- Soft Attention層の有効性を確認
  - ◆入力表現を適応的に選択し、精度を改善

今後の展望

- Soft Attention層の導入箇所の検討
  - ◆NFPに変換される前に、入力表現自体を選択する等、提案した層の導入部分にも自由度

- 適応的に選択するものの検討
  - ◆提案法自体は、データごとに適応的選択する手法のため、アルゴリズムの選択も可能