分子構造を用いた文書からの薬物相互作用抽出

Extracting Drug-Drug Interactions from Texts using Molecular Structures

浅田 真生　三輪 誠　佐々木 裕
Masaki Asada　Makoto Miwa　Yutaka Sasaki
豊田工業大学
Toyota Technological Institute
{sd17402, makoto-miwa, yutaka.sasaki}@toyota-ti.ac.jp

1 はじめに

患者に薬物を併用投与する場合，薬物の本来の作用が増強・減弱することや，副作用が起きることがあり，このことを薬物相互作用とよぶ．薬学論文等からの相互作用自動抽出を行う研究は重要であり，近年はニューラルネットワークを用いた手法 [3, 5] が注目されている．ニューラルネットワークの学習には薬学専門家により正解付けされたコーパスが必要であり，コーパス作成はコストが大きいため増やすことが難しい．そのため，既存の外部薬物データベース上から，相互作用抽出に有用な情報を低コストに獲得する研究は重要である．

また，近年，分子構造のようなグラフ構造を扱う Graph Convolutional Network (GCN) が化合物の物質推定などにおいて注目されている．

本研究では，薬物の相互作用情報と分子構造を掲載した外部薬物データベース上で，相互作用情報を教師とし，GCN によって薬物の分子構造の表現を事前学習し，獲得した分子構造表現を文書からの相互作用抽出に利用し，分子構造の利用可能性を調査することを目的とする．

2 関連研究

2.1 CNN を用いた薬物相互作用抽出

Liu らは薬物相互作用抽出の精度向上を目的とし，Convolutional Neural Network (CNN) を用いた手法 [3] を提案し，高い性能を示している．しかし，モデルの学習のために使用するコーパスは，薬学専門家によって人手で作成されたものであり，より高い性能のために正解付きコーパスを増やすことはコストが大きい．

2.2 GCN

GCN はグラフ構造を入力としたニューラルネットワークであり，分子構造を入力とした分子の物性予測などに利用されており，シミュレーションによる手法よりも高速に物性値の予測ができる．グラフ構造 G は，GCN により実数値ベクトル $\boldsymbol{g}$ に変換される．

2.2.1 Neural Fingerprint

Devenaud らは Neural Fingerprint (NFP) [1] モデルを提案し，化合物の水溶度・毒性推定において高い精度を示している．グラフ G のノード v について，v に近接するノード集合を $N(v)$ で表現する．

$$\boldsymbol{m}_v^{t+1} = \sum_{w \in N(v)} [\boldsymbol{h}_w^t; \boldsymbol{e}_{vw}] \tag{1}$$

ここで，$\boldsymbol{h}_v^t$ はノード v の更新 t ステップ目のベクトル表現であり，T ステップまで更新を行う．また，$\boldsymbol{h}_v^0$ は原子 v の入力素性である．$\boldsymbol{e}_{vw}$ はノード v, w 間のエッジの表現ベクトルである．$[\ldots;\ldots]$ はベクトルの連結を表す．

$$\boldsymbol{h}_v^{t+1} = \sigma(H_t^{deg(v)} \boldsymbol{m}_v^{t+1}) \tag{2}$$

$deg(v)$ はノード v の次数であり，$H_t^{deg(v)}$ は重み，σ はシグモイド関数である．グラフ構造を表す実数値ベクトル $\boldsymbol{g}$ は式 (3) のようになる．

$$\boldsymbol{g} = \sum_{v,t} \mathrm{softmax}(W_t \boldsymbol{h}_v^t) \tag{3}$$

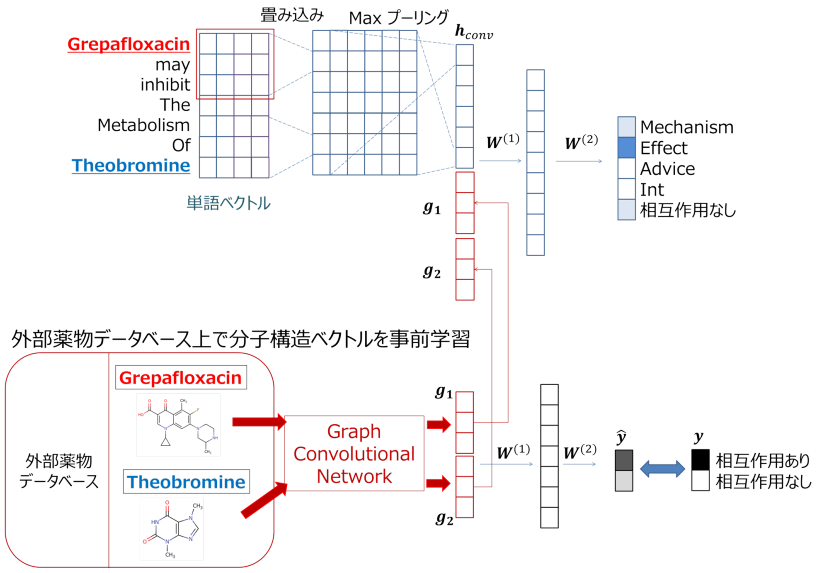

図 1: 提案手法の全体図

2.2.2 Gated-Graph Neural Network

Li らは，Gated Reccurent Unit (GRU) を用いた GCN モデルである Gated-Graph Neural Nework (GGNN) [2] を提案している．

$$\boldsymbol{m}_v^{t+1} = \sum_{w \in N(v)} A_{e_{vw}} \boldsymbol{h}_w^t \tag{4}$$

$A_{e_{vw}}$ はエッジの種類ごとに割り当てられる重みである．GRU によりノード表現ベクトルを更新する．

$$\boldsymbol{h}_v^{t+1} = \mathrm{GRU}([\boldsymbol{h}_w^t; \boldsymbol{m}_v^{t+1}]) \tag{5}$$

グラフ全体の表現ベクトル $\boldsymbol{g}$ は式 (6) のように表される．

$$\boldsymbol{g} = \sum_v \sigma(i([\boldsymbol{h}_v^T; \boldsymbol{h}_v^0])) \odot (j([\boldsymbol{h}_v^T; \boldsymbol{h}_v^0])) \tag{6}$$

ここで i, j は単層パーセプトロン，$\odot$ は全要素積である．

3 提案手法

本研究では，文書からの相互作用抽出における薬物の分子構造情報の利用可能性を検証するために，外部データベース上で分子構造表現を事前学習し，学習した分子構造表現を文書からの相互作用抽出に利用する手法を提案する．モデルの全体像を図 1 に示す．はじめに外部薬物データベース上の相互作用情報を教師とし，薬物の分子構造を入力として，分子構造のベクトル表現を事前学習する．獲得した分子構造ベクトルを正解付けされたコーパス上での CNN による相互作用抽出に利用する．

3.1 外部データベース上での分子構造ベクトルの獲得

薬物の相互作用情報と，分子構造を掲載した外部データベースを用いて，2 つの薬物の分子構造を GCN によりベクトルで表現する．それぞれの分子構造ベクトルを連結し，相互作用を持つか持たないかの二値分類問題を解くことでニューラルネットワークを学習し，その中間表現である分子構造ベクトルを獲得する．

$$\boldsymbol{g} = [\boldsymbol{g_1}; \boldsymbol{g_2}] \tag{7}$$

$$\hat{\boldsymbol{y}} = \mathrm{softmax}(\boldsymbol{W}^{(2)}\ \mathrm{relu}(\boldsymbol{W}^{(1)}\boldsymbol{g} + \boldsymbol{b}^{(1)}) + \boldsymbol{b}^{(2)}) \tag{8}$$

$\boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}$ は重み，$\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}$ はバイアスである．予測 $\hat{\boldsymbol{y}}$ と正解ラベルの交差エントロピー損失を最小にするように学習を行う．最適化には Adam を用いる．

3.2 CNN による入力文表現

薬物のペアを含む入力文 $S = (w_1, w_2, \cdots, w_N)$ が与えられたときにターゲットの薬物間の関係を予測する．単語 w_i の単語ベクトルを $\boldsymbol{w}_i$，畳み込みのウィンドウサイズの集合を $\{k_1, \cdots, k_L\}$，出力の次元数を C，重み，バイアスを $\boldsymbol{W}^{conv}$，b^{conv} とする．

$$\boldsymbol{z}_{i,l} = [(\boldsymbol{w}_{i-(k_l-1)/2}), \cdots, (\boldsymbol{w}_{i-(k_l+1)/2})] \tag{9}$$

$$m_{i,j,l} = \mathrm{relu}(\boldsymbol{W}_j^{conv} \odot \boldsymbol{z}_{i,l} + b^{conv}) \quad (10)$$

Max プーリングを行い，薬物ペアの表現を以下のように得る．

$$\boldsymbol{h}_l = [h_{1,l}, \cdots, h_{C,l}],\ h_{j,l} = \max_i m_{i,j,l}. \quad (11)$$

$$\boldsymbol{h}_{conv} = [\boldsymbol{h}_1; \ldots; \boldsymbol{h}_L]. \quad (12)$$

3.3 分子構造表現を用いた相互作用抽出

CNN の出力と，入力文中の薬物ペアの分子構造ベクトルを以下のように連結する．

$$\boldsymbol{h}_{all} = [\boldsymbol{h}_{conv}; \boldsymbol{g_1}; \boldsymbol{g_2}] \quad (13)$$

ただし，分子構造ベクトル $\boldsymbol{g_1}, \boldsymbol{g_2}$ はそれぞれ正規化を行う．薬物間の関係表現 $\hat{\boldsymbol{y}}$ を以下のように得る．

$$\hat{\boldsymbol{y}} = \mathrm{softmax}(\boldsymbol{W}^{(2)}\ \mathrm{relu}(\boldsymbol{W}^{(1)}\boldsymbol{h}_{all} + \boldsymbol{b}^{(1)}) + \boldsymbol{b}^{(2)}) \quad (14)$$

$\boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}$ は重み，$\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}$ はバイアスである．予測 $\hat{\boldsymbol{y}}$ と正解ラベルの交差エントロピー損失を最小にするように学習を行う．最適化には Adam を用いる．事前学習した単語ベクトルは fine-tuning し，分子構造ベクトルは fine-tuning しない．$\boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}, \boldsymbol{W}^{conv}$，単語ベクトル埋め込み行列に対して L2 正則化を行う．また，更新のたびに全ての学習する重みを取り出し，予測時は平均化した重みを用いる．

4 実験

4.1 実験設定

4.1.1 外部データベース上での分子構造ベクトルの事前学習

薬物データベース DrugBank に掲載された薬物のうち，相互作用を持つ薬物ペア 255,342 組を正例とし，相互作用の掲載されていない薬物ペアからランダムに正例と同じ数だけサンプリングして負例とした．正例，負例をそれぞれ 4∶1 に分割し，訓練データ，評価データとした．分子構造ベクトルの事前学習に使用したハイパーパラメータを表 1 に示す．

表 1: 分子構造ベクトル学習のハイパーパラメータ

パラメータ	値
分子構造ベクトルの次元数	50
ステップ数	4
中間層の次元数	1,000
学習率	0.001
ミニバッチサイズ	100
NFP モデルの隠れ層次元数	50
GGNN モデルの GRU 次元数	50

4.1.2 正解付きコーパス上での相互作用抽出

使用するコーパスは，Semeval-2013 Task 9 データセット [4] である．このデータセットは，薬物を含んだ文から構成され，どの単語が薬物であるかはあらかじめ特定されている．データセットは以下に示す 4 種類の相互作用が正解付けされており，本実験では薬物のペアの相互作用の有無および，相互作用を持つ場合は 4 種類のうちのどの相互作用を持つかを求める．

Mechanism: 2 つの薬物が薬力学的作用を持つ．

Effect: 2 つの薬物が薬動態学的作用を持つ．

Advice: 2 つの薬物を併用する際の推奨を表す．

Int: 相互作用を持つことのみを表す．

データセットの内訳を表 2 に示す．表 2 より，相互作用を持つペアよりも相互作用を持たないペアが多いことがわかる．コーパス中の薬物と DrugBank の薬物の対応は文字列の部分一致により判断した．コーパス中の薬物のうち DrugBank 中の薬物と対応が取れたものの割合を表 3 に示す．一致が取れなかった薬物の分子構造ベクトルの初期値は 0 ベクトルを用いた．入力文は GENIA tagger により単語分割を行った．また，Liu ら [3] と同様に，ターゲットの 2 つの薬物はそれぞれ “DRUG1”，“DRUG2” に，それ以外の薬物は “DRUGOTHER” に置き換える前処理を行った．相互

表 2: 正解付きコーパスの内訳

	訓練データ	評価データ
文数	6,976	1,299
ペア数	27,792	5,716
相互作用を持つペア数	4,021	979
相互作用を持たないペア数	23,771	4,737
Mechanism	1,319	302
Effect	1,687	360
Advice	826	221
Int	189	96

表 3: コーパス中の薬物の一致率 (%)

	DRUG1	DRUG2
訓練データ	94.51	95.71
評価データ	94.89	95.74

作用抽出の学習に使用したハイパーパラメータを表 4 に示す.

4.1.3 単語ベクトルの事前学習

単語ベクトルの事前学習は Skip-gram によって行った．事前学習に用いたコーパスは PubMed2014 のデータで，語彙数は 215,840 である．前処理によって置き換えられた “DRUG1” および “DRUG2” の単語ベクトルの初期値は “drug” と同じ値を用いた．訓練データ中に新しく出現した単語のベクトルは，事前学習した全単語の単語ベクトルの平均値を使用した．訓練データ中の単語のうち，出現頻度が 1 である単語を未知語として扱った.

5 結果

DrugBank 中の薬物ペアのそれぞれの分子構造を入力とし，相互作用あり・なしの 2 値分類を行った結果を表 5 に示す．GGNN モデルは NFP モデルよりも高い正解率を示した．続いて，SemEval 2013 Task 9 コーパスにおける相互作用抽出精度を表 6 に示す．Zheng らは LSTM と注意機構を用いた手法 [5] で，Semeval 2013 Task 9 コーパスにおいて世界最高精度を示している．NFP モデル，GGNN モデルのいずれにおいても，事前学習した分子構造ベクトルを素性に加えることにより，F 値が高くなった．また，分子構造ベクトルを加えることにより，Recall がより高くなることがわかった．Liu [3]，Zheng [5] らのモデルも分子構造を利用することで精度が上がることが期待される.

表 4: 相互作用抽出のハイパーパラメータ

パラメータ	値
単語ベクトルの次元数	200
中間層の次元数	500
畳み込みのウィンドウサイズ	{3,5,7}
畳み込みの次元数	100
学習率	0.001
ミニバッチサイズ	50
L2 正則化の係数	0.0001

表 5: DrugBank 上の 2 値分類精度

手法	正解率 (%)
NFP	94.06
GGNN	98.03

表 6: 相互作用抽出精度

手法	Precision (%)	Recall (%)	F 値 (%)
CNN	69.07	69.36	69.21
CNN+NFP	68.39	73.14	70.68
CNN+GGNN	68.98	73.14	71.00
Liu [3]	75.29	60.37	67.01
Zheng [5]	75.9	68.7	71.5

6 おわりに

あらかじめ外部薬物データベース上で学習した薬物の分子構造ベクトルを用いて CNN による薬物相互作用抽出を行い，分子構造ベクトルを用いない場合よりも精度を上げることができた．今後は外部データベース上の学習と，コーパス上の学習を同時に行う手法を検討し，分子構造のさらなる利用可能性を調査していきたい.

謝辞

本研究は JSPS 科研費 JP17K12741 の助成を受けたものである.

参考文献

[1] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pp. 2224–2232, 2015.

[2] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. International Conference on Learning Representations, 2016.

[3] Shengyu Liu, Buzhou Tang, Qingcai Chen, and Xiaolong Wang. Drug-drug interaction extraction via convolutional neural networks. *Computational and mathematical methods in medicine*, Vol. 2016, , 2016.

[4] Isabel Segura Bedmar, Paloma Martínez, and María Herrero Zazo. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). Association for Computational Linguistics, 2013.

[5] Wei Zheng, Hongfei Lin, Ling Luo, Zhehuan Zhao, Zhengguang Li, Yijia Zhang, Zhihao Yang, and Jian Wang. An attention-based effective neural model for drug-drug interactions extraction. *BMC bioinformatics*, Vol. 18, No. 1, p. 445, 2017.