

# 異種属性の内容的特徴をハイパーグラフにより統合するエンティティ表現学習

西出隆盛 三輪誠  
豊田工業大学

{sd21066,makoto-miwa}@toyota-ti.ac.jp

## 概要

エンティティの内容を表す名前や説明などのテキストで表される様々な異なる種類の属性情報に含まれるエンティティの内容的な特徴を反映したエンティティ表現を獲得する新たなハイパーグラフ表現学習手法を提案する。具体的には、属性情報をエンティティとは異なるノードとして統合したハイパーグラフを定義し、その上で新たに提案するアテンション機構を利用した異種ハイパーグラフエンコーダと属性情報からエンティティを予測するエンティティ予測損失を導入し、属性情報に含まれるエンティティの内容的な特徴をエンティティ表現に統合する。PubMed データセットを対象としたノード分類において、提案手法が既存手法と比較して高い分類性能を示すことを確認した。

## 1 はじめに

エンティティ表現の獲得は、情報検索、推薦システム、分類など多くのタスクにおいて重要である。例えば、論文エンティティには、図 1 上部のようにタイトルやアブストラクトといったテキストで表される様々な異なる種類の属性情報<sup>1)</sup>が付随している。これを反映できれば、属性情報に含まれるエンティティの内容的な特徴をより詳細に捉えた表現の獲得が期待できる。

エンティティやその属性情報のまとまりは、複数のノードをハイパーエッジと呼ばれる集合で接続できるハイパーグラフにより表現できる。近年、このようなハイパーグラフを対象としたハイパーグラフ表現学習 [1] が注目を集めている。ハイパーグラフ表現学習は、主にノード分類やリンク予測といった特定のタスクに特化した教師あり学習の枠組みで研究が行われてきた [2, 3]。さらに、事前タスクとし

1) 本論文ではエンティティに付随する様々な情報を広く属性と呼ぶ

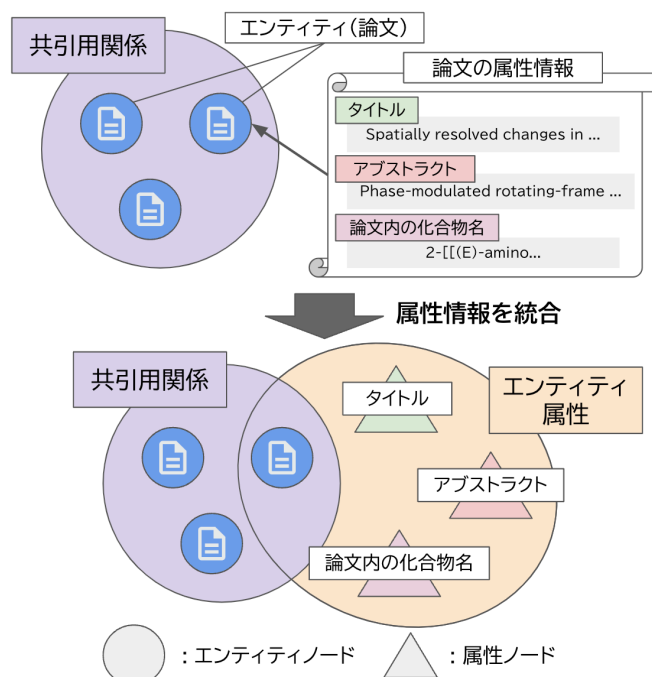


図 1 属性情報を統合したハイパーグラフの構築

てタスク非依存な汎用的表現を獲得できる自己教師あり学習が行われており、ハイパーグラフ拡張に基づく対照学習やハイパーエッジ補完などが提案されている [4, 5]。しかし、既存のハイパーグラフ表現学習手法では、エンティティと属性情報で表される異なるノードやそれらの間のハイパーエッジを区別して扱うことができていない。

図 1 下部のように、属性情報を統合したハイパーグラフでは、エンティティノードに加えて属性情報を表すノードが存在し、エンティティ間の関係だけでなくエンティティと属性ノード間の関係もハイパーエッジとして定義する。このようなハイパーグラフを扱うためには、異なる種類のノードとハイパーエッジを区別して扱い、エンティティと属性情報間の関係からエンティティの特徴を捉え、それをエンティティ表現に反映する必要がある。

そこで本研究では、こうした属性情報を統合した

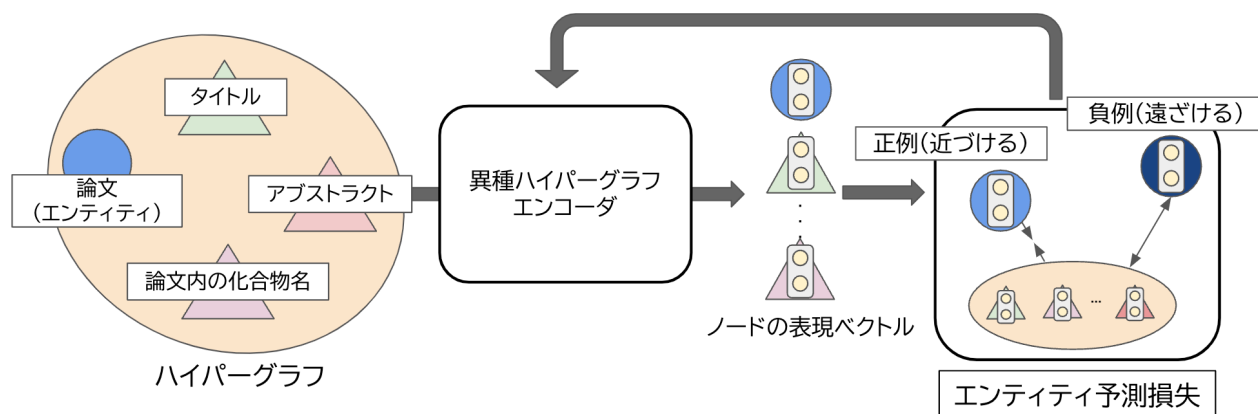


図2 提案手法の概要

ハイパーグラフにおける異なる種類のノードとハイパーエッジを考慮し、属性情報が表すエンティティの内容的特徴を反映したエンティティ表現を獲得するための新たなハイパーグラフ表現学習手法を提案する。このようなハイパーグラフに特化したエンコードのために、異なる種類のノードとハイパーエッジに対応した注意機構に基づく異種ハイパーグラフエンコーダとエンティティと属性情報間の関係性を学習するエンティティ予測損失を新たに導入する。これにより属性情報を統合したハイパーグラフにおけるエンティティの表現学習を実現し、ハイパーグラフによる構造的情報とテキストによる内容的情報の両方を考慮したエンティティ表現の獲得を目指す。

本手法の有効性の検証として、PubMed 共引用ネットワーク [6, 7] における論文の分類タスクにおける性能評価を行う。具体的には論文をノードとし、共引用関係をハイパーエッジとする基本構造に、タイトルやアブストラクトなどのテキスト属性をノードとして統合し、既存のハイパーグラフ表現学習手法と比較する。

本研究における貢献は以下である。

- エンティティとは異なるノードとして属性情報を統合したハイパーグラフを対象に、異種のノードとハイパーエッジを区別して処理する新たな異種ハイパーグラフエンコーダを提案した。
- 属性情報に含まれるエンティティの内容的特徴をエンティティ表現に反映させるために、エンティティ予測損失を導入した。
- PubMed データセットを用いたノード分類において、提案手法が既存のハイパーグラフ表現学

習手法と比較して高い分類性能を示すことを確認し、属性情報を統合する提案手法の有効性を示した。

## 2 関連研究

ハイパーグラフ表現学習は、ハイパーグラフから、その構造的・関係的特性を保持した潜在表現を学習することを目的としており、近年では、Hypergraph Neural Networks (HGNNs) [1] に基づく手法が多く提案されている。HGNNs は、ノードの特徴とハイパーグラフ構造を入力とし、ノードからハイパーエッジ、ハイパーエッジからノードへのメッセージ伝播を通じてハイパーグラフを表現空間にエンコードする。HNHN [8] はハイパーエッジ専用のエンコーダを導入し、UniGNN [3] はグラフとハイパーグラフの両方に対応する一般化されたメッセージ伝播関数を提案した。

このようなハイパーグラフ表現学習は当初、ノード分類やリンク予測といった特定のタスクに特化した教師あり学習の枠組みで研究が行われてきたが、最近では、タスク非依存な汎用的表現を獲得可能な自己教師あり学習のフレームワークが提案されている。事前タスクとして、TriCL [4] はハイパーグラフ拡張によってマルチビュー対照学習を行い、HypeBoy [5] はハイパーエッジ補完タスクと呼ばれる生成的自己教師あり学習により、汎用的表現を獲得した。

## 3 提案手法

異種のノードとハイパーエッジを区別して処理するアテンション機構を利用した異種ハイパーグラフエンコーダと、属性情報からエンティティを予測す

るエンティティ予測損失から構成される，エンティティと属性情報を表す異なるノードやそれらの間のハイパーエッジを区別して扱う新たなハイパーグラフ表現学習手法を提案する(図2)．以下，3.1節でエンコーダ，3.2節で対照損失について述べる．

### 3.1 アテンション機構を利用した異種ハイパーグラフエンコーダ

異種ハイパーグラフエンコーダでは，エンティティノードと属性ノードの異なる特性を考慮し，メッセージ伝播を繰り返すことで階層的な特徴抽出を行う．具体的には，ノードとエッジの埋め込みを用いた初期表現の獲得，ノードとハイパーエッジの双方向のメッセージ伝播，出力変換の3つの処理を行い，ノード表現を得る．

#### 3.1.1 ノードとエッジの埋め込みを用いた初期表現の獲得

エンティティノードと属性ノードの異なる特性，およびそれらの関係性を捉えるため，以下の3種類の埋め込み表現を導入する．

ノードタイプ埋め込み  $\mathbf{e}_t$ ：タイトルやアブストラクトなど，各ノードタイプの意味的特徴を表現

エンティティ属性埋め込み  $\mathbf{e}_k$ ：エンティティノードと属性ノードの異なる役割を表現

エッジタイプ埋め込み  $\mathbf{e}_r$ ：エンティティ間および，エンティティ-属性間の関係性を表現

入力特徴量  $\mathbf{X}_t$  に対し，非線形変換と埋め込みの加算により初期表現を獲得する．

$$\mathbf{H}_t^{(0)} = \text{ELU}(\text{LN}(\mathbf{W}_t \mathbf{X}_t + \mathbf{b}_t)) + \mathbf{e}_t + \mathbf{e}_k \quad (1)$$

#### 3.1.2 ノードとハイパーエッジの双方向メッセージ伝播

エッジタイプ  $r \in R$  ごとに以下の双方向のメッセージ伝播を行う．ここで， $H_e$  と  $H_n$  はアテンションヘッド数を表す．

1) ノード  $\rightarrow$  ハイパーエッジのメッセージ伝播：エッジタイプ埋め込み  $\mathbf{e}_r$  を加算した後，マルチヘッドアテンションにより特徴を集約する．

$$\mathbf{M}_{e,r}^{(l)} = \sum_{h=1}^{H_e} \text{softmax} \left( \frac{\mathbf{Q}_1^h (\mathbf{K}_1^h)^\top}{\sqrt{d_k}} \right) \mathbf{V}_1^h \quad (2)$$

2) ハイパーエッジ  $\rightarrow$  ノードのメッセージ伝播：生成されたハイパーエッジ表現からノードへ逆伝播を行う．

$$\mathbf{H}^{(l+1)} = \mathbf{H}^{(l)} + \sum_{r \in R} \sum_{h=1}^{H_n} \text{softmax} \left( \frac{\mathbf{Q}_2^h (\mathbf{K}_2^h)^\top}{\sqrt{d_k}} \right) \mathbf{V}_2^h \quad (3)$$

#### 3.1.3 出力変換

各ノードタイプごとに異なる MLP を適用し，最終的なノード表現を得る．

$$\mathbf{Z}_t = \text{MLP}_t(\mathbf{H}_t^{(L)}) \quad (4)$$

### 3.2 エンティティ予測損失

属性情報に含まれるエンティティの内容的特徴を効果的に反映させるため，エンティティと属性の関係，属性の重要度、種類を考慮したエンティティ予測損失を導入する．

#### 3.2.1 エンティティに特化した負例サンプリング

各ハイパーエッジ  $e \in E$  に対し，含まれる正例エンティティ  $v_{en} \in V_{en}^e$  以外のエンティティノードを負例として用いる．これにより，エンティティと属性情報との関係性を明示的に学習する．

#### 3.2.2 重要度を考慮した属性情報の集約

ハイパーエッジに含まれる属性ノード集合  $N_e$  の表現を集約する際，アテンション機構により各属性の重要度を考慮する．

$$\mathbf{m}_e = \sum_{v \in N_e} \alpha_v \mathbf{h}_v, \quad \alpha_v = \text{softmax}(\mathbf{w}^\top \mathbf{h}_v) \quad (5)$$

ここで， $\alpha_v$  は属性ノード  $v$  の重要度を表す．この重要度の学習により，エンティティの内容を特徴付ける重要な属性情報を選択的に集約する．

#### 3.2.3 属性の種類を考慮した対照損失

タイトルやアブストラクトなど，異なる種類の属性情報の特性を考慮するため，属性タイプに応じて適応的に調整される温度パラメータを用いた対照損失を導入する．

$$L_{\text{ours}} = -\frac{1}{|E|} \sum_{e \in E} \sum_{v_{en} \in V_{en}^e} \left( \frac{\text{sim}(\mathbf{h}_{en}, \mathbf{m}_e)}{\tau_r \tau_t} - \log \sum_{v' \in V_{en}} \exp \left( \frac{\text{sim}(\mathbf{h}_{v'}, \mathbf{m}_e)}{\tau_r \tau_t} \right) \right) \quad (6)$$

ここで， $\tau_r$  と  $\tau_t$  はそれぞれ関係タイプと属性タイプに応じた温度パラメータを表す．この損失関数により，属性情報に含まれるエンティティの内容的特徴を表現に反映する．

最終的に，モデルの学習には以下の損失関数を用いる．

$$L = L_{\text{tricl}} + \alpha L_{\text{ours}} \quad (7)$$

ここで,  $L_{tricl}$  は TriCL [4] の対照損失 (詳細は付録 B.1),  $\alpha$  は重みパラメータを表す.

## 4 実験

### 4.1 実験設定

実験には, PubMed データセット [7] を用いた. このデータセットは 3,840 件の論文ノードと 7,963 件の共引用関係から構成され, NCBI の Entrez API を通じて 3 種類の属性情報 (タイトル, 要旨, 論文の化合物名) を取得した. データの詳細な統計は付録 A.1 に示す. 各属性情報の特徴抽出には PubMedBERT [9] (出力: 768 次元) を用いた.

ベースラインとなる手法では UniGCNII [3] をエンコーダとして用い, 各手法で提案されている損失関数により学習を行った. ハイパーパラメータは付録 B.3 に示す.

評価は線形評価プロトコル [10] により行い, エンコーダの学習で得られた表現を固定して線形分類器を学習した. 比較手法として使用したベースライン手法は付録 B.2 に示す. また, 評価指標は別のシード値で 5 回実行した平均正解率と標準偏差を用いた.

### 4.2 ノード分類による評価

まず, エンティティ予測損失を用いない基本設定において, 属性情報の有効性の検証と有効な組み合わせの調査のため, 異なる属性情報の組み合わせによる比較実験を行った. 表 1 に, ノード分類タスクにおける性能の比較をまとめた. 単一属性ではタイトルが最も高い性能を示し, タイトルに要旨を加えた 2 属性の組み合わせがより高い性能を示した. これは, タイトルによる主要トピックの表現と要旨による詳細情報が補完的に働いたためと考えられる. さらに, 全ての属性情報を統合することで最高性能となり, 各属性が異なる側面から学習に寄与することを確認した.

この結果を踏まえ, 全属性を利用した上での提案手法の異なる設定とベースライン手法との比較を表 2 に示す. 実験結果から, 既存手法に対して, 提案手法が高い分類性能を示すことが確認できる. エンティティのみの結果でもベースライン手法を上回っていることから, 提案した異種ハイパーグラフエンコーダがエンティティのみの基本的なハイパーグラフ構造に対しても有効であることを示している. ま

表 1 属性情報ごとのノード分類性能 (%)

<b>単一属性</b>	
タイトル	85.01 $\pm$ 0.79
要旨	84.62 $\pm$ 0.73
化合物名	83.72 $\pm$ 0.62
<b>属性の組み合わせ</b>	
タイトル + 要旨	85.07 $\pm$ 0.65
タイトル + 化合物名	84.94 $\pm$ 0.71
要旨 + 化合物名	84.28 $\pm$ 0.76
<b>全属性</b>	<b>85.10 <math>\pm</math> 0.85</b>

表 2 手法ごとのノード分類性能 (%)

<b>ベースライン</b>	
TriCL	82.26 $\pm$ 0.47
HyperGCL	82.87 $\pm$ 0.68
HypeBoy	83.11 $\pm$ 0.63
<b>提案手法</b>	
エンティティのみ	84.29 $\pm$ 0.71
+ エンティティ予測損失	84.35 $\pm$ 0.81
エンティティ + 全属性	85.10 $\pm$ 0.85
+ エンティティ予測損失	<b>85.27 <math>\pm</math> 0.75</b>

た, エンティティのみを用いた基本設定に対し, エンティティ予測損失の追加で若干の改善が見られた. さらに, 全属性情報の統合で性能が向上し, エンティティ予測損失と組み合わせで 85.27% を達成した. これは, エンティティと属性情報を統合的に扱う本手法の有効性を示している.

## 5 おわりに

本研究では, エンティティに付随する異種属性情報の内容的特徴を反映したエンティティ表現の獲得を目的として, 属性情報を統合したハイパーグラフを対象とする新たな表現学習手法を提案した. PubMed データセットのノード分類タスクにおいて, 異なる属性情報の効果を分析した結果, 利用した全属性の統合で最も高い性能向上が確認され, 各属性が異なる側面から学習に寄与することが示された. また, 既存手法との比較において, 提案手法が高い分類性能を示すことを確認し, 属性情報の統合の有効性を実証した. 今後は提案手法の実用性と汎用性を向上させるため, より大規模なデータセットへの適用と異なる学術ドメインへの一般化に取り組む.

## 参考文献

- [1] Alessia Antelmi, et al. A survey on hypergraph representation learning. **ACM**, Vol. 56, No. 1, aug 2023.
- [2] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 33, No. 01, pp. 3558–3565, Jul. 2019.
- [3] Jing Huang and Jie Yang. Unignn: a unified framework for graph and hypergraph neural networks. In Zhi-Hua Zhou, editor, **Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21**, pp. 2563–2569. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.
- [4] Dongjin Lee and Kijung Shin. I’ m me, we’ re us, and i’ m us: Tri-directional contrastive learning on hypergraphs. In **Proceedings of the AAAI Conference on Artificial Intelligence**, 第 37 卷, pp. 8456–8464, 2023.
- [5] Sunwoo Kim, Shinhwan Kang, Fanchen Bu, Soo Yong Lee, Jaemin Yoo, and Kijung Shin. Hypeboy: Generative self-supervised representation learning on hypergraphs. In **The Twelfth International Conference on Learning Representations**, 2024.
- [6] Lise Getoor Galileo Mark Namata, Ben London and Bert Huang. Query-driven active surveying for collective classification. In **International Workshop on Mining and Learning with Graphs**, Edinburgh, Scotland, 2012.
- [7] Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Vikram Nitin, Anand Louis, and Partha Talukdar. Hypergcn: A new method for training graph convolutional networks on hypergraphs. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 32. Curran Associates, Inc., 2019.
- [8] Yihe Dong, Will Sawin, and Yoshua Bengio. Hnhn: Hypergraph networks with hyperedge neurons. **ICML Graph Representation Learning and Beyond Workshop**, 2020.
- [9] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Matthew Lungren, Tristan Naumann, and Hoifung Poon. Large-scale domain-specific pretraining for biomedical vision-language processing, 2023.
- [10] Petar Veličković, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. In **International Conference on Learning Representations**, 2019.
- [11] Tianxin Wei, Yuning You, Tianlong Chen, Yang Shen, Jingrui He, and Zhangyang Wang. Augmentations in hypergraph contrastive learning: Fabricated and generative. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, **Advances in Neural Information Processing Systems**, 2022.
- [12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.
- [13] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond, 2021.
- [14] Aaron Defazio, Xingyu Yang, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, and Ashok Cutkosky. The road less scheduled, 2024.



## A データセットの統計

### A.1 PubMed データセット

表 3 にデータセットの基本統計を示す.

表 3 PubMed データセットの統計	
ノード数	
Paper	3,840
Title	3,840
Abstract	3,840
PubChem Name	605
ハイパーエッジ数	
共引用関係	7,963
エンティティ-属性関係	3,840

線形分類器の学習実験においてはデータセットの分割比率を訓練：10%，検証：10%，テスト：80%とした.

## B 学習設定詳細

### B.1 TriCL の対照損失

エンコーダの学習で用いた TriCL の損失関数は、以下の 3 つの対照損失の重み付き和として定義される.

- ノードレベル損失：同一ハイパーエッジ内のノード間の類似性を最大化
- エッジレベル損失：共通ノードを持つハイパーエッジ間の類似性を最大化
- メンバーシップ損失：ノードとその所属ハイパーエッジ間の類似性を最大化

$$L_{tri\text{cl}} = \alpha L_{\text{node}} + \beta L_{\text{edge}} + \gamma L_{\text{mem}} \quad (8)$$

ここで、 $\alpha, \beta, \gamma$  は各損失項の重みを表すハイパーパラメータである.

### B.2 ベースライン手法

比較実験に使用した既存のハイパーグラフ表現学習手法は以下のようである.

**HyperGCL** [11]：ハイパーグラフ構造の摂動に基づく対照学習

**TriCL** [4]：3 方向の対照学習を用いたハイパーグラフ表現学習

**HypeBoy** [5]：生成的自己教師あり学習によるハ

イパーエッジ補完

### B.3 ハイパーパラメータ

#### B.3.1 異種ハイパーグラフエンコーダの設定

モデルの主要なハイパーパラメータを表 4 に示す.

表 4 異種ハイパーグラフエンコーダのパラメータ設定

パラメータ	値
埋め込み次元	512
中間層の次元	512
畳み込み層数	3
エッジアテンションヘッド数	8
ノードアテンションヘッド数	8
特徴変換層数	2
Dropout 率	0.2

#### B.3.2 ハイパーパラメータの探索

エンコーダの学習のハイパーパラメータについては、以下のパラメータの組み合わせを探索し、検証データにおける性能が最も高いものを採用した.

- エポック数: {200, 400, 600, 700, 800, 900, 1000}
- 学習率: {5e-4, 1e-3, 5e-3}
- ハイパーグラフ畳み込み層数: {1, 2, 3}
- 最適化手法: {AdamW [12], RAdam [13], AdamW (Schedule Free [14]), RAdam (Schedule Free)}

全ての属性情報とエンティティ予測損失を用いた最終実験では、以下の設定を使用した:

表 5 最終実験におけるハイパーパラメータ設定

パラメータ	値
エポック数	900
学習率	1e-3
畳み込み層数	3
最適化手法	RAdam (Schedule Free)