情報処理学会論文誌　データベース　Vol.10　No.4　36–40 (Dec. 2017)

テクニカルノート

制限付き識別ランダムウォークによるグラフベースのラベル拡張

木村 正成[1,a)]　若林 啓[2,b)]

受付日 2017年6月10日, 採録日 2017年8月7日

概要：多くの半教師あり学習の目標は，ラベル付きデータとラベルなしデータをうまく組み合わせて分類性能の高いモデルを作ることである．半教師あり学習でよく用いられる手法の 1 つにラベル伝播があるが，確信度の低いデータにもラベルをつけてしまうため性能が落ちてしまうという問題がある．本研究では，有向グラフ上でのランダムウォークにいくつかのルールを課した新しい手法を提案し，深層学習などのより強力な分類器を学習することによって，ラベル付きデータが極端に少ないケースであっても高い性能のモデルを学習することができる．実験では，提案手法をベンチマークデータに適用し，既存の単純なラベル伝播を行ってラベルを増やした手法と比較してそれを上回る結果が得られた．

キーワード：グラフ構造，ランダムウォーク，ラベル拡張，半教師あり学習

Graph-based Label Extension Using Restricted Discriminative-random Walk

Masanari Kimura[1,a)]　Kei Wakabayashi[2,b)]

Received: June 10, 2017, Accepted: August 7, 2017

Abstract: The goal of many semi supervised learning is to make a model with high classification performance by successfully combining labeled data and unlabeled data. Label propagation is one of the techniques often used in semi supervised learning, but there is a problem that performance is lowered because labels are attached to data with low confidence. In this research, we propose a new method that imposes several rules on random walk on directed graph, learning stronger classifiers such as deep learning, and so on, it is a case with extremely few labeled data even with a high performance model you can learn. Experiments show that the proposed method is applied to benchmark data, and compared with the method in which labels are increased by performing existing simple label propagation, the result exceeding that is obtained.

Keywords: graph structure, random walk, label extension, semi supervised learning

1. はじめに

データ工学において，ベクトルデータのカテゴリ分類に帰着できる問題は多く存在する．近年では，機械学習を用いることにより，大量のデータに対する分類も自動的に行うことが可能になっているが，このような分類器を構築するためには，一般に大量のラベル付きデータが必要となる．しかし，人手で大量のデータにラベルをつけるのは非常にコストのかかる作業である．一方，ラベルなしデータであれば，インターネットやカメラ映像，センサデータなど様々なドメインにおいて自動で大量に入手できる．このことを背景として，ラベル付きデータとラベルなしデータの両方を活用して学習器を構築する半教師あり学習の手法が研究されている [1], [2].

半教師あり学習の手法にはいくつか種類があるが，本研究では，データの類似度に基づいて構築したグラフ構造を利用する手法に注目する．グラフ構造における半教師あり

1 筑波大学情報学群知識情報・図書館学類
College of Knowledge and Library Sciences, School of Informatics, University of Tsukuba, Tsukuba, Ibaraki 305–8550, Japan
2 筑波大学図書館情報メディア系
Faculty of Library, Information and Media Science, University of Tsukuba, Tsukuba, Ibaraki 305–8550, Japan
a) s1411518@u.tsukuba.ac.jp
b) kwakaba@slis.tsukuba.ac.jp

学習の手法のうち，最も一般的なものの 1 つにラベル伝播がある [3]．これは，グラフ上で距離の近いノードは同じラベルを持つ可能性が高いという仮定に基づいて，すでにラベルが付与されているノードからラベルが付与されていないノードにラベルを伝播させる手法である．ラベル伝播は，ラベルなしデータの分布を考慮した分類を行えることから，ラベル付きデータが非常に少ない場合に有効である．一方で，ある程度の量のラベル付きデータが利用可能である場合には，類似度グラフの構築を行わずに SVM やニューラルネットワークなどのベクトル分類器を適用した方が，ベクトル空間上の情報を直接利用できるため有利である場合が多い．

本研究では，ラベル伝播によってすべてのデータにラベルを付与するのではなく，確信度の高い一部のデータにのみ付与する “ラベル拡張” を行い，増量したラベル付きデータを用いてベクトル分類器を学習するアプローチを検討する．この目的のため，ランダムウォークとして解釈できるラベル伝播の手法 [8] において，その動作に一定のルールを導入した制限付き識別ランダムウォークを提案する．これを用いたラベル拡張を行い，拡張したデータを教師情報として深層学習などの教師あり分類器の学習を行うことで，従来の教師あり学習の手法よりも高い精度で分類を行うことができることを示す．

2. 関連研究

ラベル伝播に関する研究は数多く存在する [4], [5]．その多くは，グラフ上で近くに位置するデータは同じラベルを有する可能性が高いという仮定に基づく [6]．

Rosenfeld ら [5] は，ラベルが伝播する過程を感染モデルとしてとらえ，各ノードに，自身がそのラベルをつけられるかどうかにかかわらず隣接ノードにラベルを伝播させる潜伏期間を定めることによって，ラベル伝播を試みている．しかし，この手法では一度ラベルが伝播されたノードはそれ以降変更の余地がないものとしている．このため，ノイズの影響を受けやすく，精度が落ちてしまう可能性がある．また，Cohen の研究 [7] では，距離拡散に基づいた有向グラフ上での半教師あり学習の手法を提案している．この手法ではラベル付けされていない各ノードを起点として学習を試みているが，ラベル付けされていないノードからラベル付けされたノードへの距離を求める計算コストは大きくなりやすいという問題がある．一方で，ラベル付けされたノードからラベル付けされていないノードへの距離は効率的に計算が可能である場合が多い．

これらの手法は，すべてのデータに対してラベルを付与する半教師あり学習器の構築を目的としている．本研究では，これらの研究で得られた知見をベクトルデータのラベル拡張に援用し，ベクトル分類器の精度向上を目指す．

3. ランダムウォークによるラベル伝播

本研究のランダムウォークの動作の元となったラベル伝播の手法に，Callut ら [8] によって提案された Discriminative Random Walks がある．以降，この手法を D-Walks と表記する．重み付きグラフ $G = (V, E, W)$ が与えられる．ここで，ノード集合を V，エッジ集合を E，重み集合を W とする．また，$|V|$ はノード数，$|E|$ はエッジ数とする．A は G の $|V| \times |V|$ の隣接行列で，$a_{q_1 q_2}$ はエッジ $q_1 \to q_2$ 間の重みを表す．ただし，q_1 と q_2 はそれぞれ V 中のノードとする．

グラフ G の一部のノードのみにラベルが付与されているものとする．ラベル付きノード集合を $L \subseteq V$，ラベルなしノード集合を $U \subseteq V \backslash L$ とする．ラベルの集合を Y とし，ノード $q \in L$ のラベルを $y_q \in Y$，クラス y に属するノード集合を $L_y = \{q | y_q = y\}$，L_y の要素数を $|L_y|$ と表す．

時刻 t においてノード q から $q^{'}$ に遷移する確率は，エッジの重み $a_{qq'}$ に依存して以下のように定義される．

$$P[X_{t+1} = q^{'} | X_t = q] = P_{qq'} = \frac{a_{qq'}}{\sum_{q' \in N} a_{qq'}} \qquad (1)$$

D-Walks では，最大歩数 L をパラメータとして，以下のルールに従うランダムウォークを考える．

- ラベル y の付与されたノードから開始し，同じラベル y がつけられたノードに到達するか，最大歩数 L を超えた場合終了する．
- 開始ノードと同じラベル y の付与されたノード（開始ノードでもよい）に到達したウォークを，試行成功とする．
- 最大歩数 L までに同じラベルがつけられたノードに到達しなかったウォークを，試行失敗とする．
- 試行の際に別のラベルがつけられたノードを経由してもよい．

D-Walks は，ラベル y についてこのようなランダムウォークを無限回行ったときに，試行成功したウォークがノード q を通過する期待回数に基づいて，q のラベルの予測を行う．

これは，形式的には以下の条件を満たすウォークの集合を考えることに対応する．開始ノードと終了ノードがいずれも L_y に含まれていて，かつ経由ノードがいずれも L_y に含まれていないような長さ l のウォークの集合を D_l^y とする．また，この条件を満たすような最大歩数 L 以下のすべてのウォークの集合を $D_{\leq L}^y = \bigcup_{l=1}^{L} D_l^y$ とする．式 (1) の遷移確率に基づいたウォーク d の生起確率を $p(d)$ とし，ウォーク d においてノード q を通過した回数を $n_d(q)$ とする．ノード q がクラス y についてのランダムウォーク中に訪れられる回数の期待値を表す関数を $B_L(q, y)$ と表記すると，これは以下のように定義できる．

$$B_L(q,y) = E[n_d(q)|D^y_{\leq L}] = \sum_{d \in D^y_{\leq L}} p(d) n_d(q) \qquad (2)$$

D-Walks では，ラベル付けされていないノード q は，最も通過回数の期待値が大きいクラスに割り当てられる．

$$\hat{y} = \arg\max_{y \in Y} B_L(q,y) \qquad (3)$$

D-Walks の特徴は，ランダムウォークの動きに一定のルールを決め，最大歩数に上限を決めることでグラフ上のコミュニティ間の違いを明確にし，クラスの予測を効率的に計算できる点である．一方で，付与されるラベルの期待値が複数のラベルで拮抗している場合でも分類を行ってしまっているため，グラフ上のコミュニティが散在しているケースでは精度が落ちてしまう．

4. ランダムウォークによるラベル拡張

本研究では，すべてのノードにラベルを付与する必要のあるラベル伝播とは異なり，確信度の高いノードにのみラベルを付与するラベル拡張の手法を提案する．また，以下では無向グラフではなく有向グラフを使用することで，コミュニティ構造をより反映できることを示す．与えられるグラフは，ベクトルデータの集合から構成される類似度グラフとする．**図 1** に有向グラフ上で各ノードから 2 近傍にエッジを張った際のランダムウォークの動作を示す．ここで，1 つのノードは 1 つのベクトルデータに対応し，各ノード間の重み $a_{qq'}$ はベクトルのコサイン類似度によって与えられるものとする．また，図中の実線矢印は有向エッジ，点線矢印はランダムウォークの動作を意味する．図 1 において，ノード 1 がラベル付きノードでその他のノードがラベルなしノードとする．ノード 1 からランダムウォークを開始し，ノード 1 に戻ってくることができた経路に含まれるノードのみに新しくラベルを付与する．ノード 2 およびノード 3 はノード 1 に帰ってくることができるため，ノード 1 と同じラベルを得られる．一方そのほかのノードからは，ノード 1 に向かう経路がないため，ラベルが付与されない．また，**図 2** に，コミュニティをまたがって同じラベルが与えられているケースを示す．同じラベルが付与された別のコミュニティ間に，もう 1 つコミュニティがある場合，無向グラフを使用した場合ではコミュニティを横断する際にラベルが伝播してしまう．しかし有向グラフを用いることで，破線で囲まれている部分グラフとその他の部分グラフがそれぞれ分割されているものと考えられる．

4.1 Restricted Discriminative Random Walks

上記をふまえて，本節では D-Walks に新しく制約を加え，ラベル拡張を行うためのランダムウォークの手法を提案する．以下，提案手法を RD-Walks と表記する．RD-Walks では，試行の際に別のラベルがつけられたノードの経由を

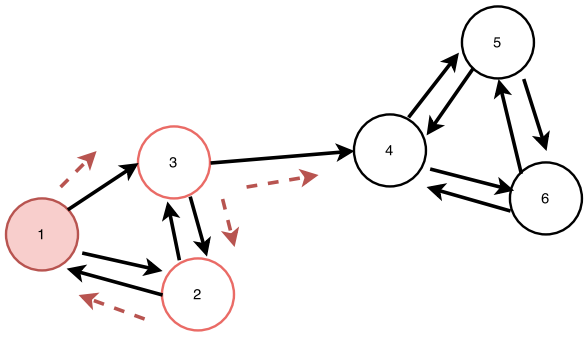


図 1　有向グラフ上で各ノードから 2 近傍にエッジを張った際のランダムウォークの動作

Fig. 1　The behavior of random walk on 2 neighborhood directed graph.

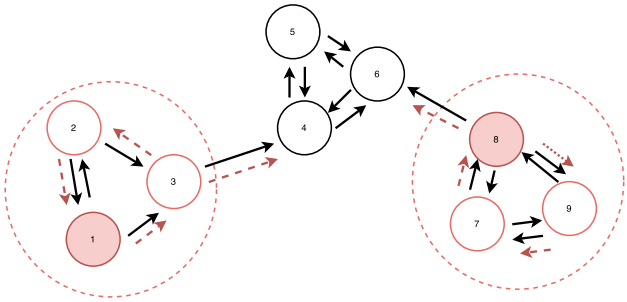


図 2　コミュニティをまたがって同じラベルが与えられている場合の Restricted D-Walks の動作

Fig. 2　The behavior of the restricted discriminative random walk when the same label is given to nodes in separate communities.

禁止する．また，別のラベルがつけられたノードに到達した時点でランダムウォークを打ち切るようにする．さらに，新たなパラメータとして最大歩行距離 L_d を導入する．ランダムウォークの際に経由したノード間の重みの総和がこの値を超えたとき，ランダムウォークを終了するようにする．これらのルールの上でランダムウォークが試行成功したとき，経由したノードにラベルを割り振る．これにより，すべてのノードにラベルを割り振ることができなくなるが，割り振ることができたラベルは確信度がより高いといえる．ここで，ランダムウォークが試行成功するとは，上記のルールのもとで開始ノードと同じラベルが付与されているノードに到達した状態をいう．

つまり，ランダムウォークに制約を課し，かつ有向グラフを用いることで，あるコミュニティのラベルが他のコミュニティに流れ込むことを防ぎ，より精度の高いラベル拡張が可能となる．

RD-Walks のパラメータとして，最大歩数を L_s，最大歩行距離を L_d を所与とする．重み付き有向グラフ $G = (V, E, W)$ が与えられる．開始ノードと終了ノードがいずれも L_y に含まれていて，かつ経由ノードがいずれも L_y に含まれていないような，長さ l，合計のエッジ重み w のウォークの集合を $D^y_{l,w}$ とする．長さが最大歩数 L_s 以

下で，かつ合計エッジ重みが最大歩行距離 L_d 以下のすべてのウォークの集合を $D^y_{\leq L_s,L_d} = \bigcup_{l\leq L_s} \bigcup_{w\leq L_d} D^y_{l,w}$ とする．これを用いて，式 (2) を修正した以下の式によって，クラス y のランダムウォークにおける q の期待訪問回数を定義する．

$$B_{L_sL_d}(q,y) = E[n_d(q)|D^y_{\leq L_s,L_d}] = \sum_{d\in D^y_{\leq L_s,L_d}} p(d)n_d(q) \tag{4}$$

RD-Walks では，ノード q には以下の式 (5) によって付与ラベルを決定する．

$$\hat{y_q} = \begin{cases} -1 & (\max(B_{L_sL_d}(q,y)) = 0) \\ \arg\max_{y\in Y} B_{L_sL_d}(q,y) & (otherwise) \end{cases} \tag{5}$$

ただし，$\hat{y_q}$ が -1 のとき，ノード q にラベルは付与されないものとする．$\hat{y_q}$ が -1 となるのは，すべてのラベル付きノードからランダムウォークを行っても，ノード q を 1 度も経由しない場合である．

4.2　RD-Walks によるラベル拡張

上記で説明した RD-Walks を複数回繰り返すことによって，ラベル拡張を行う．ここで，N 回目の RD-Walks は，$N-1$ 回目でラベル y が付与されたノードを，ラベル付きノード集合 L_y に加えて実行する．RD-Walks の試行回数を増やすごとに，ラベルの拡張範囲は増加するが，拡張されたラベルの精度も低下していく．対象とするデータによって試行回数，最大歩長および最大歩行距離を調節することによって，ラベルの拡張範囲と精度のトレードオフを自由に制御できる．実験では，RD-Walks の試行回数は 5 回程度で良い性能が得られた．

5.　実験

提案手法と従来手法に対するラベル拡張の精度と拡張したラベルを適用した教師あり学習の性能の比較実験を行う．提案手法は，各データに対してコサイン距離をもとに 20 近傍にエッジを張った有向グラフ上でのラベル拡張を行い，新たに獲得したラベル付きデータを用いて教師あり学習器を学習する．最大ステップ数は 5，最大歩行距離は各データから張られているすべてのエッジの平均としている．

5.1　実験データおよび実験環境

実験では，MNIST 手書き数字データ[*1]，UCI の Mushroom データセット[*2]，Iris データセット[*3]，および CORA データセット[*4]を使用する．これらのデータセットを学習用データとテスト用データに分割する．この学習用データから無作為に 1 クラスあたり数データを抽出してラベル付きデータとし，その他の学習用データをラベルなしデータとする．評価実験はテスト用データに対して行う．また，CORA データセットについては，全データ中から 1 クラスあたり 10 データを抽出してラベルを付与してラベル拡張を行った後，全データに対して分類を行い正解率を求めている．**表 1** に各データセットの構成を示す．

[*1] http://yann.lecun.com/exdb/mnist/
[*2] https://archive.ics.uci.edu/ml/datasets/mushroom
[*3] https://archive.ics.uci.edu/ml/datasets/iris
[*4] https://relational.fit.cvut.cz/dataset/CORA

表 1　各データセットの構成（∗ ではテスト用データと学習用データに含まれる）

Table 1　Statistics of the dataset. (* indicates that it is included in test data and learning data.)

データセット	ラベル付データ数	ラベル無データ数	テストデータ数	クラス数
MNIST	100	50000	10000	10
Mushroom	20	5686	2438	2
Iris	30	100	50	3
CORA	70	2708	∗	7

表 2　各データセットにおける各手法の比較結果（∗ 部分は Rosenfeld ら [5] の論文の図 2 を参照した）

Table 2　Comparison result of each method in each dataset. (* referred to Fig. 2 in Rosenfeld et al. [5].)

	MNIST	Mushroom	Iris	CORA
CNN	0.727	0.559	0.660	
MLP	0.707	0.724	0.960	
linear-SVM	0.693	0.605	0.940	
poly-SVM	0.594	0.517	0.820	
k-NN	0.670	0.751	0.960	
labelProp	0.098	0.528	0.667	
InfProp				0.480 ∗
提案手法+CNN	**0.909**	0.632	0.711	0.213
提案手法+MLP	0.823	**0.761**	**0.980**	**0.556**

5.2　比較方法

比較する手法は，提案手法を用いて拡張したラベル付きデータを用いて学習した多層パーセプトロン（以下 MLP）と畳み込みニューラルネットワーク（以下 CNN），ラベル拡張をせずにもともと与えられたラベル付きデータのみで学習した CNN と MLP，線形カーネルおよび多項式カーネルを用いたサポートベクタマシン，k 近傍法，Zhu ら [3] によるラベル伝播法，Rosenfeld ら [5] によるラベル伝播手法の 10 種類について行う．MLP は 3 層の全結合層からなり，CNN は畳み込み層が 2 層，プーリング層が 1 層，全結合層が 1 層のものを使用する．また，提案手法を用いる場合と用いない場合で，ニューラルネットワークは共通のものを使用する．評価指標は，予測結果全体と正解データがどれだけ一致しているかを判断する正解率を用いる．

5.3　実験結果

表 2 に各データセットに対して 1 クラスあたり 10 デー

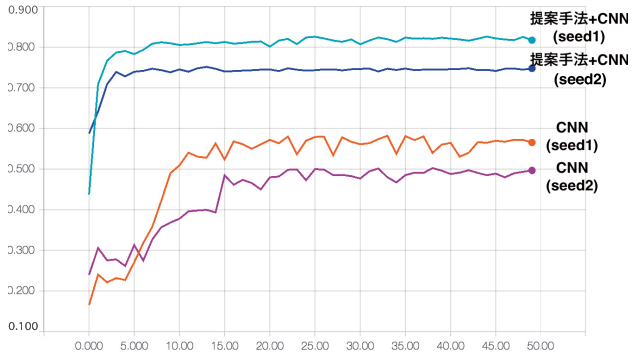


図 3　MNIST データセットに対して，1 クラスあたり 2 枚のラベルを与えて学習させた際の各 epoch での予測結果（横軸が学習回数，縦軸が正解率）

Fig. 3　Results of learning by giving two labels per class to the MNIST dataset. (The number of learning is plotted on the horizontal axis and the accuracy rate on the vertical axis.)

タにラベルを付与した場合の各手法の実験結果を示す．提案手法を用いたラベル拡張を行わなかった場合に比べて，行った場合の方が良い結果が得られていることが分かる．また，本研究の提案手法である RD-Walks は既存の様々な教師あり，および半教師あり学習の手法の事前学習器として機能するため，状況に応じてそれらを組み合わせることができる．今回の実験では，すべてのデータセットに対して教師あり学習器は共通のものを使用したが，各データセットに適した学習器と組み合わせることによってより分類性能の高いモデルを学習できる．

また，**図 3** に MNIST データセットにおいて 1 クラスあたり 2 データのみにラベルを付与した場合のラベル拡張を行った場合と行っていない場合の実験結果を示す．モデルは両方とも CNN を使用している．提案手法を用いた場合と用いていない場合の両方とも，最初に選択するラベル付きデータによって結果が左右されやすい一方で，シードによっては，ラベル付きデータが極端に少ない場合でも 8 割以上の正解率が得られていることが分かる．

6.　結論

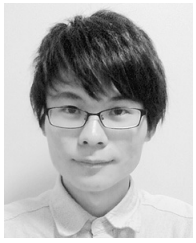
本稿では，グラフ上のランダムウォークを用いて，ラベル拡張を行い，拡張したデータを新たに教師情報として教師あり学習器の学習を試みた．実験結果から，従来手法と比較して提案手法が高い正解率でのラベル拡張が可能であることを示した．

グラフベースの半教師あり学習手法全般にいえることだが，学習の性能はグラフ構造に強く依存する．Wauquier ら [9] や Yan ら [10] のように，グラフの構築に注目した学習を取り入れ，より効率的なグラフの構築を行うところまで含めた研究などが今後の課題である．

謝辞　本研究の一部は，JSPS 科研費（課題番号 16H02904）の助成によって行われた．

参考文献

[1] Zhu, X. and Goldberg, A.B.: Introduction to semi-supervised learning, *Synthesis lectures on artificial intelligence and machine learning*, 3.1, pp.1–130 (2009).
[2] Rosenberg, C., Hebert, M. and Schneiderman, H.: Semi-supervised self-training of object detection models, *7th IEEE Workshop on Applications of Computer Vision* (2005).
[3] Zhu, X. and Ghahramani, Z.: Learning from labeled and unlabeled data with label propagation, Technical Report CMU-CALD-02-107, 1 (2002).
[4] Liu, W., Wang, J. and Chang, S.-F.: Robust and scalable graph-based semisupervised learning, *Proc. IEEE*, 100.9, pp.2624–2638 (2012).
[5] Rosenfeld, N. and Globerson, A.: Semi-Supervised Learning with Competitive Infection Models, *arXiv preprint*, arXiv:1703.06426 (2017).
[6] Wang, F. and Zhang, C.: Label propagation through linear neighborhoods, *IEEE Trans. Knowledge and Data Engineering*, 20.1, pp.55–67 (2008).
[7] Cohen, E.: Semi-supervised learning on graphs through reach and distance diffusion, *arXiv preprint*, arXiv:1603.09064 (2016).
[8] Callut, J., Françoisse, K., Saerens, M. and Dupont, P.: Classification in graphs using discriminative random walks, *MLG* (2008).
[9] Wauquier, P. and Keller, M.: Metric learning approach for graph-based label propagation, *arXiv preprint*, arXiv:1511.05789 (2015).
[10] Yan, S., Xu, D., Zhang, B., Zhang, H.-J., Yang, Q. and Lin, S.: Graph embedding and extensions: A general framework for dimensionality reduction, *IEEE trans. pattern analysis and machine intelligence*, 29.1 (2007).



木村 正成

筑波大学情報学群知識情報図書館学類在学中．機械学習の研究に従事．

若林 啓　（正会員）

2012 年法政大学大学院博士課程修了．博士（工学）．同年筑波大学図書館情報メディア系助教．機械学習の研究に従事．日本データベース学会，ACM 各会員．

（担当編集委員　上土井 陽子）