

node2vecと機械学習による属性推定

推定対象の属性: 性別, 教育形態

教育形態	内容
フォーマル教育	通常の学校教育 (Ex. 大学で講義を受けて単位を取る など形式的な教育)
インフォーマル教育	人々との触れ合いを通して知識, 技術を獲得する 学校外の教育 (Ex. 講演会, ハッカソン)
ノンフォーマル教育	学校外の教育ではあるが形式的な教育 (Ex. 放送大学, 通信教育など)

使用データ

Facebookのソーシャルグラフの交友関係

ソーシャルグラフの基本情報

ノード数	4039
リンク数	88234

交友関係の例

236 122

236 65

236 43

237 90

237 123

237 459

ユーザID 236は
ユーザID 122と
交友関係がある

使用データ

- 性別のデータの概要

元データが匿名化されていたため

以降のラベルの内容はデータ数から推測

内訳

ラベル 0 (男性)	2504
ラベル 1 (女性)	1531

使用データ

- ・教育形態のデータの概要

教育形態1: 内訳

ラベル 0 (教育を受けていない)	1236
ラベル 1 (教育を受けた)	2806

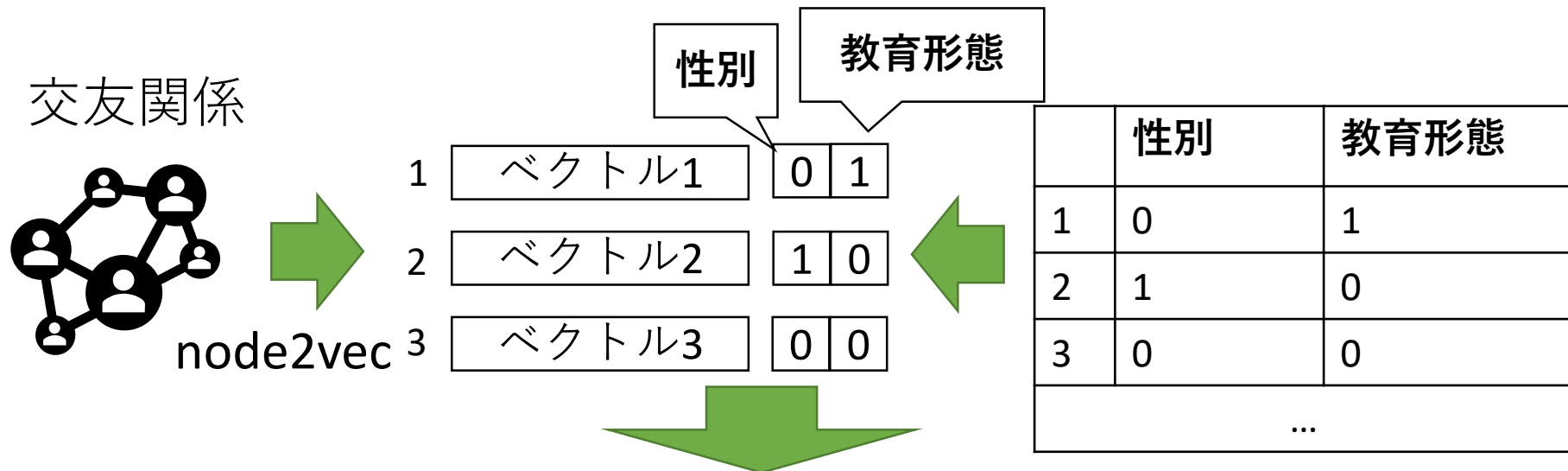
教育形態2: 内訳

ラベル 0 (教育を受けていない)	2831
ラベル 1 (教育を受けた)	1221

教育形態3: 内訳

ラベル 0 (教育を受けていない)	1499
ラベル 1 (教育を受けた)	2543

入出力と処理方式



XGBoost or ロジスティック回帰

識別結果

学習データ: テストデータ = 8:2 に分割
学習データ: **BalancedBagging or アンダーサンプリング**
テストデータ: **アンダーサンプリング**

評価指標

- 正答率 (Accuracy)
- 適合率 (Precision)
- 再現率 (Recall)
- F値
- 混同行列

結果: 性別

- 学習データ, テストデータ: アンダーサンプリング
ロジスティック回帰により識別

	Accuracy	Precision	Recall	F値
ラベル0	0.620	0.63	0.60	0.61
ラベル1		0.61	0.64	0.63

- 混同行列

	ラベル0(予測)	ラベル1(予測)
ラベル0(真実)	177	119
ラベル1(真実)	106	190

結果: 教育形態1

- 学習データ: BalancedBagging テストデータ: アンダーサンプリング
XGBoostにより識別

	Accuracy	Precision	Recall	F値
ラベル0	0.610	0.63	0.52	0.57
ラベル1		0.59	0.70	0.64

- 混同行列

	ラベル0(予測)	ラベル1(予測)
ラベル0(真実)	130	120
ラベル1(真実)	75	175

結果: 教育形態2

- 学習データ: BalancedBagging テストデータ: アンダーサンプリング
XGBoostにより識別

	Accuracy	Precision	Recall	F値
ラベル0	0.714	0.69	0.77	0.73
ラベル1		0.74	0.65	0.70

- 混同行列

	ラベル0(予測)	ラベル1(予測)
ラベル0(真実)	197	58
ラベル1(真実)	88	167

結果: 教育形態3

- 学習データ: BalancedBagging テストデータ: アンダーサンプリング
XGBoostにより識別

	Accuracy	Precision	Recall	F値
ラベル0	0.624	0.63	0.62	0.62
ラベル1		0.62	0.63	0.63

- 混同行列

	ラベル0(予測)	ラベル1(予測)
ラベル0(真実)	184	114
ラベル1(真実)	110	188

考察: 性別

- 交友関係のみでは性別が識別されにくかった理由


Facebookでは、実際に深い交友関係がなくとも同じコミュニティ(会社の同僚など)という理由でユーザーをフォローしているケースが多い

➡ 同じコミュニティならば性別に関係なくフォローしている交友関係のグラフのみでは交友関係の深さが分からない

➡ 交友関係**のみ**では識別され**にくい**

考察：教育形態

- ・ 教育形態2が識別されやすかった理由
 - ・ **教育形態2に属するノードは交友関係に違いがある**
 - ・ インフォーマル教育：
フォーマル教育， ノンフォーマル教育と教育環境が異なる

 **教育形態2 = インフォーマル教育** と推測

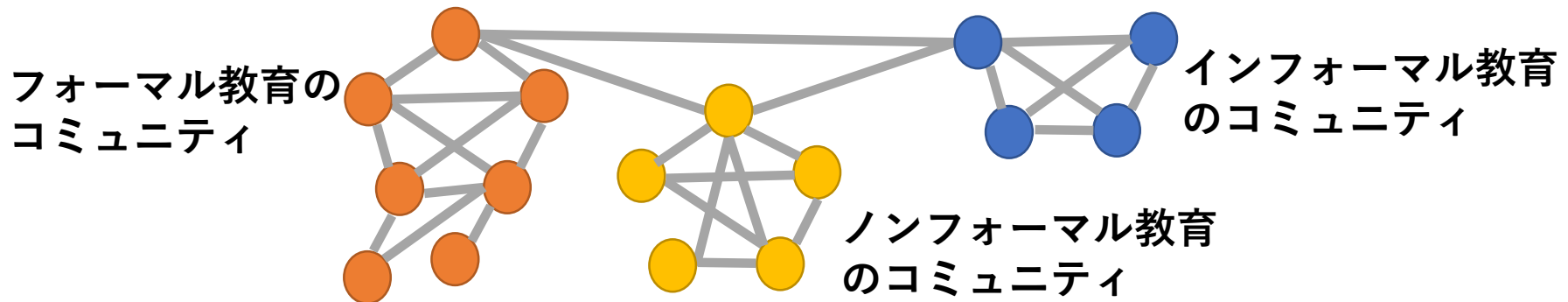
- ・ 教育形態1, 3(フォーマル教育， ノンフォーマル教育)では交友関係から属性推定をすることは困難

考察: 教育形態

- ・ インフォーマル教育が識別されやすかった理由
収集データはアメリカが対象

➡ アメリカでは通常、フォーマル教育
インフォーマル教育は珍しい

Facebookでは経歴(学歴, 職歴)を登録可能
経歴から「知り合いかも」によりフォロー可能



インフォーマル教育を受けた少数のユーザー間で
形成されたコミュニティがあったため識別されやすかった