

1. 概要
2. 実験方法
3. 今後の予定

卒業研究進捗報告

nanimonaiyo

山本 藤也 (Touya Yamamoto)
u220067@st.pu-toyama.ac.jp

富山県立大学 情報システム工学科

October 10, 2025

目的

- 学生の履修選択を支援するシステムの開発
- キャリアパスに基づいた科目推薦の仕組みの追加

進捗

- 提案手法と他の手法の比較、実験方法の検討

手法の説明

1. 概要
2. 実験方法
3. 今後の予定

- ある事象（例：「IT 業界への就職」）が起こるか否かを、複数の要因（各科目の成績）から予測するための、統計的な回帰モデルの一種。各科目がキャリアパスに与える影響の強さを「重要度スコア」として算出する。

特徴

- 最大の特徴:L1 正則化という機能により、学習の過程で関連性の低い科目の重要度スコアを自動的にゼロにする。これにより、多数の科目の中から、特定のキャリアパスにとって本当に重要な「キー科目」だけを効率的に選び出すことができる。

1. 概要
2. 実験方法
3. 今後の予定

手法の説明

- 多数の「決定木」というシンプルな予測モデルを作り、それらの多数決によって最終的な予測を行う、非常に強力なアンサンブル学習モデル。

比較内容

- L1 正則化が各科目の影響を個別に（）見るのでに対し、ランダムフォレストは「『A』と『B』を両方履修している」といった、科目の複雑な組み合わせ（非線形な関係）を捉えるのが得意。

選んだ理由

- 一般的に非常に高い予測精度を誇るため、提案手法の性能を測るために強力なベンチマークとなる。また、同様に科目の重要度を算出できるため、「キー科目の発見能力」という観点でも直接比較が可能。

1. 概要
2. 実験方法
3. 今後の予定

手法の説明

- データ点を複数のグループ（例：「IT 業界に進んだ組」と「それ以外」）に分けるための最も明確な境界線を、数学的に見つけ出すモデル。

比較内容

- 同じ線形モデルだが、境界線の引き方が異なる。ロジスティック回帰は全データを考慮するのに対し、SVM は境界線に最も近いデータ（サポートベクター）を重視する。

選んだ理由

- 提案手法とは異なるアプローチで線形の分類を行う代表的な手法であり、異なる考え方でデータを分析した場合の性能差を見ることで、より深い考察が可能になる。

1. 概要
2. 実験方法
3. 今後の予定

手法の説明

- 「この科目の成績が良い学生は、IT 業界に進む確率が高い」といった、確率をベースに予測を行う、非常にシンプルで高速なモデル。

比較内容

- 「全ての科目は、互いに独立している（影響を与え合わない）」という単純な（ナイーブな）仮定に基づいて成り立っている。これにより、複雑な計算をせずに高速な予測が可能。

選んだ理由

- 複雑なモデルの性能を評価する際に、比較の基準となるシンプルなベースラインとしてよく用いられる。この単純な手法と比較することで、提案手法に最低より優る価値があるかを確認する。

手法の説明

- ランダムフォレストと同様に決定木を多数作るが、前の木が間違えた部分を次の木が重点的に学習する、という「改善を繰り返す」アプローチを取る手法。

比較内容

- ランダムフォレスト以上に、科目の複雑な組み合わせから精度の高い予測を行う能力に長けている。

選んだ理由

- 多くのデータ分析コンペで勝利を収めているモデルである。この最高性能モデルと比較することで、提案手法の実用的な性能レベルを客観的に示すことができる。

1. 概要
2. 実験方法
3. 今後の予定

予測精度 (量的評価)

- 手法：十分割交差検証で実験し、正解率、適合率、再現率、F1で評価
- 目的：どのモデルが最も正確にキャリアパスを予想できるか

推薦内容の妥当性 (質的評価)

- 手法：在校生全員への推薦を作成し、推薦頻出科目を集計
- 目的：どのモデルが、より妥当なキー科目を推薦するか

推薦の安定性 (頑健性評価)

- 手法：学習データをサンプリングしながら推薦を繰り返し、結果の一貫性を Jaccard 係数でスコア化
- 目的：モデルの推薦は、学習データが少し変わっても安定しているか

1. 概要
2. 実験方法
3. 今後の予定

客観的な有効性検証のため、意図的に特定の統計的傾向を持つ合成データセットを生成した。※詳細省略

生成内容

- 科目とキャリアの間に、あらかじめ定義された「関連度（履修確率）」を埋め込む。（例：IT 業界に進む学生は、ソフトウェア工学を 80% の確率で履修）
- これにより、モデルが「正解」を正しく学習できるかを検証可能。

推薦内容の妥当性（質的評価）

- データ規模 (3 段階): 卒業生 150 人 / 600 人 / 3000 人
- 関連度の強さ (2 段階): 閾値 0.8 (強い相関) / 0.5 (弱い相関)
- ノイズ量 (3 段階): 10% / 20% / 30%

1. 概要
2. 実験方法
3. 今後の予定

データセット条件	L1正則化	ランダムフォレスト	SVM	ナイーブベイズ	勾配ブースティング
少量データ (n=150, t=0.8, n=10%)	0.2203	0.2509	0.1973	0.2365	0.2334
中量データ (n=600, t=0.8, n=10%)	0.2538	0.3088	0.2268	0.2279	0.3201
大量データ (n=3000, t=0.8, n=10%)	0.2819	0.3377	0.273	0.2418	0.3367
中量・低関連度 (n=600, t=0.5, n=10%)	0.2356	0.2763	0.2455	0.1751	0.3037
中量・高ノイズ (n=600, t=0.8, n=30%)	0.1051	0.0679	0.1006	0.0956	0.0575

結果

- 純粹な予測精度ではやはり複雑な組み合わせが得意なランダムフォレストと勾配ブースティングが優秀であったが、提案した手法はほかの線形的分析よりは優秀な結果を残した。
- また、ノイズが多いテストデータでは L1 正規化が優れた性能を残した。

推薦内容の妥当性

11/13

- 概要
- 実験方法
- 今後の予定

モデル名	1位	回数	2位	回数	3位	回数	4位	回数	5位	回数
L1正則化ロジスティック回帰	ソフトウェア工学	35	電気電子材料学	31	キャリア形成論	28	データベース論	9	ネットワーク設計論	8
ランダムフォレスト	ソフトウェア工学	35	電気電子材料学	28	キャリア形成論	27	データベース論	9	アルゴリズムとデータ構造	8
SVM (Linear)	ソフトウェア工学	35	電気電子材料学	29	キャリア形成論	27	データベース論	9	ネットワーク設計論	8
ナイーブベイズ	ソフトウェア工学	29	電気電子材料学	28	キャリア形成論	27	ネットワーク設計論	8	経済学 I	8
勾配ブースティング (LightGBM)	ソフトウェア工学	33	電気電子材料学	28	キャリア形成論	27	データベース論	9	ネットワーク設計論	8

モデル名	1位	回数	2位	回数	3位	回数	4位	回数	5位	回数
L1正則化ロジスティック回帰	電気電子材料学	34	ソフトウェア工学	34	キャリア形成論	31	CAD/CAM	10	ネットワーク設計論	8
ランダムフォレスト	電気電子材料学	34	キャリア形成論	28	ソフトウェア工学	27	CAD/CAM	10	経済学 I	8
SVM (Linear)	電気電子材料学	34	ソフトウェア工学	34	キャリア形成論	31	CAD/CAM	10	ネットワーク設計論	8
ナイーブベイズ	電気電子材料学	29	ソフトウェア工学	28	キャリア形成論	27	ネットワーク設計論	8	確立システム	8
勾配ブースティング (LightGBM)	電気電子材料学	33	ソフトウェア工学	29	キャリア形成論	27	CAD/CAM	10	確立システム	8

結果

- 上が IT 業界への推薦、下がメーカー業界への推薦
- 今回はデータ数 600 のテストデータで実験を行ったが、IT 業界への推薦でキー科目でない科目が全ての手法において 2 位に入っていたが、別のテストデータで試したら大きく結果が異なったため、テストデータの影響が大きく実験方法を再検討する必要があると考える。

1. 概要
2. 実験方法
3. 今後の予定

データセット条件	L1正則化	ランダムフォレスト	SVM	ナイーブベイズ	勾配ブースティング
少量データ (n=150, t=0.8, n=10%)	0.9571	0.6855	0.8631	0.8897	0.7822
中量データ (n=600, t=0.8, n=10%)	1	1	1	1	1
大量データ (n=3000, t=0.8, n=10%)	1	1	1	1	1

結果

- データ数が少ない 150 の時を除いて常に安定していたため、この実験に関してもあまり有効でないと考える。

今後の予定

13/13

課題

- 実験方法の見直し
- 中間発表のポスター準備

1. 概要
2. 実験方法
3. 今後の予定