

はじめに

データ分析の種類

協調フィルタリ
ング

進捗状況

2.2 教学データ
から得られる知見

おわりに

教学データからのモチベーション向上・キャリアパス支援のための情報推薦機構

平松 楓也

富山県立大学 情報基盤工学講座

July 8, 2020

背景

近年の就職活動は売り手市場と言われていたが、コロナウイルスの影響で世界的に経済状況が悪化しており、買い手市場に推移していく可能性が考えられる。また、大手企業へ就職を考えた場合、応募人数が多く狭き門であることが多いため、企業は企業がより求めている人材を採用すると思われる。そのため、学生の間にも、より効率的に企業が求める人材になるための勉強が必要になると思われる。

目的

過去の卒業生の就職先や、学業成績、野外活動のデータをクラスタリングし、在校生がより効率的に就職活動を行えるよう対話型の情報推薦機構の基礎技術を開発する。

はじめに

データ分析の種類

協調フィルタリ
ング

進捗状況

2.2 教学データ
から得られる知見

おわりに



入学したばかりの大学1年生、卒業後就職したい企業が決まっている学生
その企業に就職するためにはどうすれば効率的に動けるかわからない



協調フィルタリングで過去の卒業生の
データからやるべきことを推薦



やるべき勉強、活動が明確になりモチベーション向上
就職したかった企業への内定

説明

事実を説明する
見つける

例

どんな人が何を買っているか？
ある広告がどれだけ売りに貢献しているか？

手法

BI、クラスタリング、アソシエーション分析

説明

未来や欠測値を予想する

例

ある商品群を閲覧した人の性別は？
広告を出稿したらどれだけ売り上げが上がるのか？

手法

分類・回帰、統計的機械学習、協調フィルタリング

説明

最適解を探す

例

利益を最大化するための、最適な仕入れ量は？
売上を最大化するには、どこに広告を出稿すべきか？

手法

最適化、実験結果

協調フィルタリングとは、Amazon が開発したレコメンドエンジンで、多くのユーザの嗜好情報を蓄積し、あるユーザと嗜好の類似した他のユーザの情報を用いて自動的に推論を行う方法論である。

また、協調フィルタリングには二種類あり、ユーザベース協調フィルタリングとアイテムベース協調フィルタリングがある。

ユーザベース協調フィルタリングでは「ユーザ A は未評価アイテム I に対して、当該ユーザと似たような嗜好をしている他ユーザと同じような評価をするだろう」という仮定に基づいている。

ユーザベース協調フィルタリング

履歴から
類似ユーザ
を見つける

| | 商品A | 商品B | 商品C | 商品D |
|------|-----|-----|-----|-----|
| ユーザA | ○ | - | ○ | ○ |
| ユーザB | × | ○ | - | × |
| ユーザC | ○ | ○ | × | - |
| ユーザD | ○ | × | ○ | ? |

類似ユーザAはDを評価高くしているのでおすすめできそう

アイテムベース協調フィルタリングでは「アイテム同士の類似度とあるユーザ A の過去に評価したアイテムの評価点を用いて未評価アイテム I の評価点を予測する」というアプローチである。

アイテム間協調フィルタリング

似た評価の
商品を見つける。
商品Aと商品Dは
似た人に
買われやすい

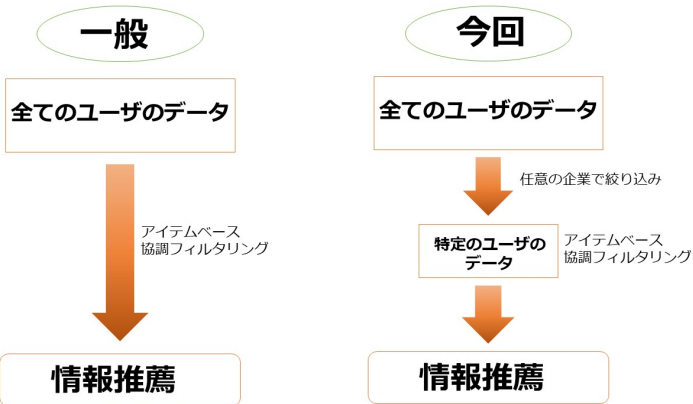
| | ユーザA | ユーザB | ユーザC | ユーザD |
|-----|------|------|------|------|
| 商品A | ○ | - | ○ | ○ |
| 商品B | × | ○ | - | × |
| 商品C | ○ | ○ | × | - |
| 商品D | ○ | × | ○ | ? |

商品Aの購買者評価と似た評価の商品Dを勧めよう

今後行うアイテムベース協調フィルタリングについて

10/23

一般に使われる協調フィルタリングは全ユーザのデータを基にフィルタリングを行うのに対し、今回は、ユーザ A が就職を希望している企業に就職したユーザのみでフィルタリングを行い情報推薦を行おうと考えている。



はじめに

データ分析の種類

協調フィルタリング

進捗状況

2.2 教学データ
から得られる知見

おわりに

1 章 はじめに

2 章 教学データの活用（従来研究）

2.1、教学データの構成・内容（論文 5 本）

2.2、教学データから得られる知見（論文 5 本）

3 章 （従来研究）（数学的な手法）

3.1

3.2 情報推薦の仕組み

4 章 提案手法（オリジナリティ）

5 章 数値実験並びに考察

6 章 まとめと今後の課題

- ・ csv ファイルを処理，管理しやすいように複数に分けることにした．

- ・ アドミッション__出身高校・入試種別.csv
- ・ アドミッション__受験科目・成績.csv
- ・ カリキュラム__履修・評価.csv
- ・ カリキュラム__科目情報.csv
- ・ キャリア関連__インターンシップ.csv
- ・ キャリア関連__就職情報.csv
- ・ キャリア関連__資格・免許.csv
- ・ 課外活動__サークル.csv
- ・ 課外活動__アルバイト.csv
- ・ 卒業後__企業側.csv
- ・ 卒業後__学生.csv

はじめに

データ分析の種類

協調フィルタリ
ング

進捗状況

2.2 教学データ
から得られる知見

おわりに

| | A | B | C | D | E |
|----|----------|------|------|---|---|
| 1 | StudentN | 出身高校 | 入試種別 | | |
| 2 | 1713001 | 宇都宮中 | 前期 | | |
| 3 | 1713002 | 四日市西 | 前期 | | |
| 4 | 1713003 | 菟道 | 前期 | | |
| 5 | 1713004 | 片山学園 | 後期 | | |
| 6 | 1713005 | 恵那 | 前期 | | |
| 7 | 1713006 | 藤枝明誠 | 推薦 | | |
| 8 | 1713007 | 氷見 | 前期 | | |
| 9 | 1713008 | 四日市西 | 後期 | | |
| 10 | 1713009 | 松阪 | 前期 | | |

図 1: 出身高校・入試種別

| | A | B | C | D | E |
|----|----------|------|------|---|---|
| 1 | StudentN | センター | 二次試験 | | |
| 2 | 1713001 | 390 | 277 | | |
| 3 | 1713002 | 389 | 292 | | |
| 4 | 1713003 | 400 | 281 | | |
| 5 | 1713004 | 874 | 0 | | |
| 6 | 1713005 | 389 | 281 | | |
| 7 | 1713006 | 0 | 0 | | |
| 8 | 1713007 | 395 | 254 | | |
| 9 | 1713008 | 885 | 0 | | |
| 10 | 1713009 | 388 | 283 | | |

図 2: 受験科目・成績

・出身高校のリストは県大の主な出身校を載せたサイトがあったのでそれを参考にした・点数は平均点を富山県立大学の H29 の平均点を参考に標準偏差 10 でランダムに発生

| | A | B | C | D | E | |
|----|----------|-------|-------|-------|-------|---|
| 1 | StudentN | 経済学 1 | 富山と日本 | 環境論 1 | 環境論 2 | E |
| 2 | 1713001 | 1 | 2 | 1 | 5 | |
| 3 | 1713002 | 3 | 5 | 2 | 3 | |
| 4 | 1713003 | 5 | 5 | 5 | 2 | |
| 5 | 1713004 | 2 | 3 | 4 | 5 | |
| 6 | 1713005 | 2 | 3 | 5 | 1 | |
| 7 | 1713006 | 4 | 3 | 4 | 4 | |
| 8 | 1713007 | 3 | | 3 | 5 | |
| 9 | 1713008 | 2 | 5 | 2 | 2 | |
| 10 | 1713009 | 5 | 3 | 5 | 4 | |

図 3: 履修・評価

| | A | B | C | D | E |
|----|----------|----------|----------|---|---|
| 1 | StudentN | 学部卒業後就職先 | | | |
| 2 | 1713001 | 就職 | デンソーテクノ | | |
| 3 | 1713002 | 就職 | スギノマシン | | |
| 4 | 1713003 | 就職 | コマツNTC | | |
| 5 | 1713004 | 就職 | 富山村田製作所 | | |
| 6 | 1713005 | 進学 | YKK AP | | |
| 7 | 1713006 | 進学 | スズキ | | |
| 8 | 1713007 | 就職 | 立山科学グループ | | |
| 9 | 1713008 | 進学 | 豊田合成 | | |
| 10 | 1713009 | 進学 | 澁谷工業 | | |

図 4: 就職情報

| | A | B | C | D | E | F | G | H |
|----|----------|-----------|-----------|----------|-----------|------|---------|---|
| 1 | StudentN | 普通自動車運転免許 | TOEIC公開試験 | MOS Word | MOS Excel | 簿記2級 | 基本情報技術者 | |
| 2 | 1713001 | 1 | 0 | 0 | 0 | 0 | 1 | |
| 3 | 1713002 | 1 | 0 | 0 | 0 | 0 | 0 | |
| 4 | 1713003 | 1 | 0 | 0 | 0 | 0 | 0 | |
| 5 | 1713004 | 1 | 0 | 0 | 0 | 0 | 0 | |
| 6 | 1713005 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 7 | 1713006 | 1 | 0 | 0 | 0 | 0 | 0 | |
| 8 | 1713007 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 9 | 1713008 | 1 | 0 | 0 | 0 | 0 | 0 | |
| 10 | 1713009 | 1 | 0 | 0 | 0 | 0 | 0 | |

図 5: 資格・免許

・就職先は富山県立大学の HP を参考にし、学科、学部、院ごとに別ランダム振り分け

・0 が未取得、1 が取得済み

| | A | B | C | D | E |
|----|----------|------------|---|---|---|
| 1 | StudentN | サークル | | | |
| 2 | 1713001 | Chat Box | | | |
| 3 | 1713002 | 無所属 | | | |
| 4 | 1713003 | 無所属 | | | |
| 5 | 1713004 | アカペラサークル | | | |
| 6 | 1713005 | フットサルサークル | | | |
| 7 | 1713006 | PDC | | | |
| 8 | 1713007 | 軟式野球部 | | | |
| 9 | 1713008 | 無所属 | | | |
| 10 | 1713009 | TRPG・映画研究会 | | | |

図 6: サークル

| | A | B | C | D | E |
|----|----------|-------------|---|---|---|
| 1 | StudentN | アルバイト | | | |
| 2 | 1713001 | 串網 | | | |
| 3 | 1713002 | アルビス | | | |
| 4 | 1713003 | 秋吉 | | | |
| 5 | 1713004 | アルバイトなし | | | |
| 6 | 1713005 | 個別指導塾スタンダード | | | |
| 7 | 1713006 | 串網 | | | |
| 8 | 1713007 | アルバイトなし | | | |
| 9 | 1713008 | アルバイトなし | | | |
| 10 | 1713009 | セブンイレブン | | | |

図 7: アルバイト

・サークルも県大に実在するサークルを全て同じ確率でランダムに振り分け、所属率は 50 %

・アルバイト先は県大周辺にあるアルバイト募集している店舗をランダムで振り分け、アルバイトしている確率は 90 %

- ・クラスター分析とは、異なる性質のものが混ざり合った集団から、互いに似た性質を持つものを集め、クラスターを作る方法である。
- ・クラスター分析には、大きく分けて2種類あり、階層クラスター分析と非階層クラスター分析に分けられる

【階層クラスター分析】

- ① 結合過程が理解しやすい
- ② クラスター数の決定は分析後

【非階層クラスター分析】

- ① 大量サンプルであってもクラスターの分類が可能
- ② クラスター数の決定は分析前

今回は階層クラスター分析を行う

【環境】

- ① jupyter notebook
- ② これは python 言語の延長線上にあるもので，データ分析により向いている．
- ③ anaconda からインストール可能で，分割して実行結果を表示できる点が一番のメリット



【得たい知見】

- ① センター試験と二次試験の各科目の結果から、その人が進学か就職するのか分類したい。
- ② このほかにもクラスター分析したら面白そうなものを考えたい

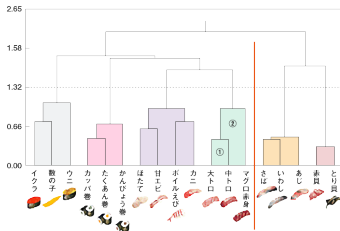


図 9: 階層クラスター分析の例

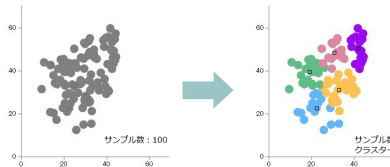


図 10: 非階層クラスター分析の例

はじめに

データ分析の種類

協調フィルタリ
ング

進捗状況

2.2 教学データ
から得られる知見

おわりに

【お膳立て】

- ① 県大のセンター試験科目の国語、数学、理科、外国語と二次試験の数学と理科の点数を使う。
- ② 就職か進学は入試の合計点数によって決めており、ある値より高いと進学になる。
- ③ 今回は要素数が大量にならないよう電子・情報科のみで分析を行う。

| | 国語セ | 数学セ | 理科セ | 外国語セ | 数学二次 | 理科二次 |
|-----|-----|-----|-----|------|------|------|
| 133 | 57 | 135 | 66 | 167 | 138 | 132 |
| 134 | 0 | 0 | 0 | 0 | 0 | 0 |
| 135 | 59 | 123 | 75 | 156 | 156 | 105 |
| 136 | 61 | 119 | 63 | 159 | 157 | 133 |
| 137 | 36 | 118 | 58 | 151 | 137 | 140 |

図 11: 使ったデータの一部

【変数設定】

- ① クラスター間の距離の計算 (method) は単純平均距離 (average) にした
- ② 距離計量 (metric) はデフォルトのユークリッド距離を使う

【その他の method と metric】

- ① method には最短距離、最長距離、単純平均距離、加重平均距離がある
- ② metric にはユークリッド距離、チェビシェフ距離、市街地距離、コサイン類似度などがある

図 12: クラスター分析結果

まとめ

- ① 本来進学か就職かを分類したかったが、入試種別の分類になってしまっている
- ② 前期、後期、推薦で点数の差が大きいのが問題だと思うので、正規化するなどして差を縮める必要があったかもしれない
- ③ 一応、正しく分類出来ていることは確認出来た

今後の課題

- ① クラスター分析を使って面白い結果が得られそうなことを考える必要があるそう
- ② 卒論 2.2 ではもっと他の手法で得られる知見もやっていきたいが教学 IR の分野だけで調べてもあまりなかったのでその他の分野で自分の研究に応用できそうなものを探していきたい。また、自分でできるか進めていきたい