

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

制約ボルツマンマ  
シン

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

August 26, 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、学生の間には、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

制約ボルツマンマ  
シン

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上げに貢献しているか？

## 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るべきか？

## 手法

最適化、実験結果

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

制約ボルツマンマ  
シン

おわりに

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

10/24

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。

**一般**

全てのユーザのデータ

アイテムベース  
協調フィルタリング

**情報推薦**

**今回**

全てのユーザのデータ

任意の企業で絞り込み

特定のユーザの  
データ

アイテムベース  
協調フィルタリング

**情報推薦**

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

制約ボルツマンマ  
シン

おわりに

1 章 はじめに

2 章 QOL の向上のための仕組み（従来研究）

2.1

2.2 モチベーションに繋がるコーチング（やり方、方法）

2.3 情報推薦の仕組み（ガイド、導く）

3 章 教学データの利活用（従来研究）

3.1 教学データの構成・内容（論文 5 本）

3.2 情報推薦の仕組み

3.3 データマイニングにおける評価基準（ジニ不純度とか） 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

制約ボルツマンマ  
シン

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

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

制約ボルツマンマ  
シン

おわりに

	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

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

制約ボルツマンマ  
シン

おわりに

	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			

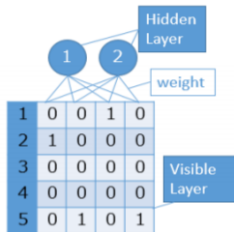
図 7: アルバイト

・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 %



制約ボルツマンマシン (Restricted Boltzmann Machine : RBM) とは、入力となる可視層 (visible layer) と隠れ層 (hidden layer) の二層構造のニューラルネットワーク構造を持ち、可視ノード間、隠れノード間には接続がない (条件付き独立性を持つ) ボルツマンマシンである。また、協調フィルタリングに適用する場合は可視変数  $v$  を拡張して考える。可視層を最大評価値の分だけ増やし、選んだ一つに 1 をつけ他は 0 をつける。



協調フィルタリングにおける制約ボルツマンマシンの応用

Application of Restricted Boltzmann Machines to  
Collaborative Filtering

図 8: 拡張した制約ボルツマンマシンの例

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

制約ボルツマンマ  
シン

おわりに

可視層バイナリ値ベクトルを  $\mathbf{v}$ , 隠れ層バイナリ値ベクトルを  $\mathbf{h}$ , 可視層バイアス項を  $\mathbf{a}$ , 隠れ層バイアス項を  $\mathbf{b}$ , エッジに対応した重みベクトルを  $\mathbf{w}$  である. また, これかのパラメータをまとめて  $\theta = (\mathbf{b}, \mathbf{c}, \mathbf{w})$  ボルツマン分布は以下の式で定義される.

$$p(\mathbf{v}, \mathbf{h} \mid \theta) = \frac{1}{Z(\theta)} \exp(-\Phi(\mathbf{v}, \mathbf{h}, \theta)) \quad (1)$$

また, エネルギー関数  $\Phi(\mathbf{v}, \mathbf{h}, \theta)$ , 正規化項  $Z(\theta)$  を以下の式で定義される.

$$\Phi(\mathbf{v}, \mathbf{h}, \theta) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_i \sum_j w_{ij} v_i h_j \quad (2)$$

$$Z(\theta) = \sum_{\mathbf{v}} \sum_{\mathbf{h}} \exp(-\Phi(\mathbf{v}, \mathbf{h}, \theta)) \quad (3)$$

可視変数  $\mathbf{v}$  が与えられたときの隠れ変数  $\mathbf{h}$  の条件付き分布  $p(\mathbf{h}|\mathbf{v}, \theta)$  を求める.

$$\begin{aligned} p(\mathbf{h} | \mathbf{v}, \theta) &= \frac{p(\mathbf{v}, \mathbf{h} | \theta)}{p(\mathbf{v} | \theta)} \\ &= \frac{p(\mathbf{v}, \mathbf{h} | \theta)}{\sum_{\mathbf{h}} p(\mathbf{v}, \mathbf{h} | \theta)} \end{aligned} \quad (4)$$

さらに計算すると

$$p(\mathbf{h} | \mathbf{v}, \theta) = \prod_j \frac{\exp\left(h_j(b_j + \sum_i w_{ij}v_i)\right)}{1 + \exp\left(b_j + \sum_i w_{ij}v_i\right)} \quad (5)$$

RBM は条件付き独立性をもつので可視層が固定されているときの、隠れ層の条件付確率は以下の式で表される

$$\begin{aligned} p(h_j = 1 \mid \mathbf{v}, \boldsymbol{\theta}) &= \frac{\exp(b_j + \sum_i w_{ij} v_i)}{1 + \exp(b_j + \sum_i w_{ij} v_i)} \\ &= \text{sigmoid} \left( b_j + \sum_i w_{ij} v_i \right) \end{aligned} \quad (6)$$

同様に隠れ変数  $\mathbf{h}$  が与えられたときの可視変数  $\mathbf{v}$  の条件付き分布  $p(\mathbf{v} \mid \mathbf{h}, \boldsymbol{\theta})$  が得られる。

$$\begin{aligned} p(v_i = 1 \mid \mathbf{h}, \boldsymbol{\theta}) &= \frac{\exp(a_i + \sum_j w_{ij} h_j)}{1 + \exp(a_i + \sum_j w_{ij} h_j)} \\ &= \text{sigmoid} \left( a_i + \sum_j w_{ij} h_j \right) \end{aligned} \quad (7)$$

RBM の学習とは、与えられた訓練データをもっともうまく説明しそうなパラメータをもつモデルを作成することである。可視層にデータを入れると、隠れ層のフィルタを通して再び可視層に値が返ってくる。その結果を入力と比較し最適な出力が得られる方向にパラメータを調整する。しかし、このパラメータ更新の計算をする際に組み合わせ爆発の問題があるため、何らかのサンプリングを行いその平均で期待値を近似する必要がある。

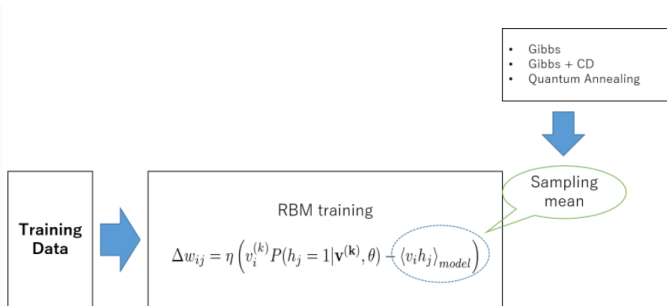


図 9: RBM の学習

## Contrastive Divergence(CD) 法

ギブスサンプリングより処理を早くしたもので、サンプリングには MCMC(マルコフ連鎖モンテカルロ法) を用いる。

### 処理手順

- ① 可視層の初期状態として、入力データ  $\mathbf{v}_0$  を用いる。
- ② 隠れ層の各ユニットの値を、 $p(\mathbf{h}_j|\mathbf{v}, \theta)$  の確率の二項分布からサンプリングする。
- ③ (2) で得た隠れ層の状態から、次ステップの可視層の値を  $p(\mathbf{v}_i|\mathbf{h}, \theta)$  の確率の二項分布からサンプリングする。
- ④ これを 1 回だけ行い、(3) で得た可視層の状態を用いて各期待値を計算する。

はじめに

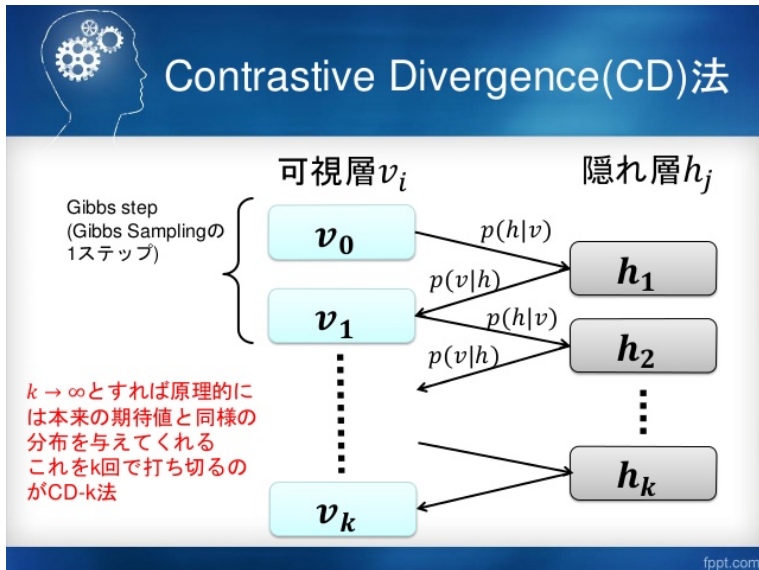
データ分析の種類

協調フィルタリ  
ング

進捗状況

制約ボルツマンマ  
シン

おわりに



## まとめ

- ① 制約ボルツマンマシンの協調フィルタリングへの応用に関する論文と記事が少ない
- ② 中国の人が `python` で制約ボルツマンマシンを使った協調フィルタリングのコードを見つけた.
- ③ 現在コードを読んでいる最中で, 購入した本も参考にすれば `python` のプログラムの作成はできると思う.