PWScup2024 team11

群馬大学情報学部　Gunmataro117（岡嶋佳歩, 熊谷紫恩, 江口誠, 田中聖也, 中曽根真衣, 中島崚杜, 矢島琴恵）

1.匿名化手法

【方針】
クロス集計表で値が可能な限り保持されるように、ランダマイズを行う

【加工方法】
① 10000行のレコードの中からランダムで1行選ぶ
② その行から更にランダムに1セル選ぶ
③ 下図の加工ルールに従って値を変更する
④ ①～③を全てのレコードについて1つ以上のセルが変更されるまで繰り返す

nステップ目 B11.csv

2	56	247	260	653	673	…
	3					

0,1,2,4,5のどれかに変更

n+1ステップ目 B11.csv

2	56	247	260	653	673	…
	0					

【加工ルール】
・nステップ目での集計表がオリジナルの集計表の値に近づくようにratingの値を選択する
・”ZIP-code”など基本属性には重みづけをする

“ZIP-code”×”56”の集計表の例

オリジナルの集計表

	0	1	2	3	4	5
6	10	10	10	10	10	10
10						
12						
…						

n回目の集計表

	0	1	2	3	4	5
6	7	9	15	12	11	9
10						
12						
…						

n+1回目の集計表

	0	1	2	3	4	5
6	8	9	15	11	11	9
10						
12						
…						

【メリット】
・集計表を使った分析に有用
・それぞれのセルの変更は独立してランダマイズされるため，レコード内のセル同士の関連を断ち切れる

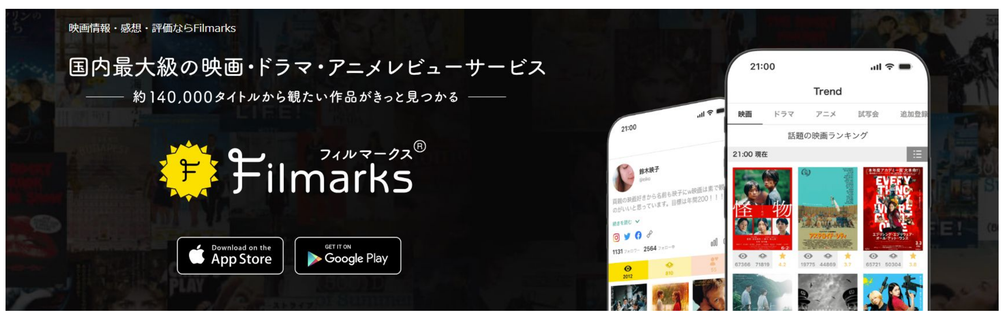
【改善点】
・重みづけの値を最適化する必要はある
　今回はセンシティブなデータ（ZIP-code -> Occupation -> Age）に
　それぞれ重みづけを行った.
　→ センシティブなデータがよりクロス集計表の値が保持される

ポイント
・クロス集計表が全部で1219個できる
　→ 全ての集計表を見ながら、
　46カラム × 10000レコード = 46万セルを変えていくので3~7日程度かかる
　チームメンバー全員各自のPCでプログラムを走らせた

2.攻撃手法

【DB再構築攻撃 方針】
最頻値と、filmarksというレビューサイトのデータを用いた推定を組み合わせて行った
　※ チーム別に異なった手法を適用したが、今回は一例を紹介する

【filmarksのデータ】
filmarks上での映画それぞれに対しての評価をそのまままとめた。
現実の攻撃ではこうしたレビューサイトの情報も活用するかと考えた。



【方法】
① Bxx_b、Cxx_0~9について、各映画に対するratingの最頻値を出す
② 各映画について①の二つがどちらも0ならば黒塗りの箇所は0
③ それ以外の箇所は、下図のオレンジ色の値を元に推定する

Bxx_b.csv

2	56	247	260	653	673	…
					*	
		*				
				*		
	*					
⋮	⋮	⋮	⋮	⋮	⋮	⋮

各人のratingの傾向(平均値)

→	1.56	→ 3
→	1.33	→ 2
→	2.02	→ 4
→	2.51	→ 5
→	0.87	→ 1

値の置き換え

最頻値　0 0 1 1 0 0

A.csvでの各人のratingの傾向(平均値)を6段階に分割したもの

rating	置き換える値
0.15~0.59	0
0.59~1.03	1
1.03~1.47	2
1.47~1.91	3
1.91~2.35	4
2.35~2.79	5

filmarksでの各映画のrating

映画ID	2	56	247	260	653	673	…
評価	3.7	発見できず	3.6	3.9	3.4	3.4	…

【計算例(赤枠)】
(1 + 4 + 3.6) / 3 = 2.8666　→ 3と推定

3.実際に分析に利用

↓「Age」と「ZIP-code」の組合せごとの、各映画を高評価(4と5)した人の度数

匿名加工前

Animation(B)	1	18	25	35	45	50	56
0	142	324	331	177	253	183	104
100	115	240	294	210	228	147	109
200	142	282	314	218	245	170	93
300	153	295	352	229	244	177	129
400	131	353	335	232	243	162	105
500	125	218	242	184	182	117	107
600	75	181	213	131	148	76	41
700	98	235	246	165	190	92	72
800	74	237	229	152	187	127	85
900	179	386	516	334	349	222	135

匿名加工後

Animation(C)	1	18	25	35	45	50	56
0	150	346	361	210	241	161	85
100	117	238	332	185	244	154	124
200	113	338	350	238	241	157	89
300	131	316	412	230	248	170	138
400	131	321	415	278	243	152	109
500	149	231	267	174	216	121	86
600	76	139	220	131	160	86	51
700	92	251	254	171	188	97	88
800	69	249	268	165	170	123	92
900	197	405	574	307	317	208	140

割合(匿名加工後データ)

Animation	1	18	25	35	45	50	56
0	0.29	0.3	0.29	0.29	0.27	0.26	0.24
100	0.28	0.27	0.28	0.25	0.3	0.32	0.3
200	0.25	0.31	0.28	0.3	0.28	0.29	0.25
300	0.26	0.29	0.3	0.29	0.28	0.31	0.31
400	0.28	0.26	0.29	0.29	0.26	0.27	0.28
500	0.34	0.3	0.28	0.29	0.33	0.34	0.28
600	0.32	0.23	0.3	0.27	0.29	0.31	0.28
700	0.26	0.31	0.27	0.29	0.29	0.28	0.29
800	0.27	0.29	0.29	0.28	0.27	0.3	0.28
900	0.32	0.29	0.29	0.29	0.27	0.31	0.28

←Animation映画を高評価した人の割合

分析結果
・度数が多いわりに年齢区分が25と35の人にはアニメは人気ではない
・ZIP-codeが「500」の人には比較的アニメが高評価の傾向がある

匿名加工データの有用性について
　匿名加工後のクロス集計表【「Age」と「ZIP-code」の組合せと各映画(46作品)の高評価(4と5)の維持が確認できた。

分析結果利用方法(ユースケース)
・地域
　-地域活性化(聖地巡礼, サイン会のスケジュール)
　-映画館の再上映計画
・年齢
　-地域の年齢分布からマーケティング
　　(高齢化地域にはこんな映画を...など)
　-50代に人気 → 健康意識向上プログラムとのコラボ　50代に人気のキャラを使って...
　　(例：運動、食生活)
　-映画の上映時間の検討

・観客の属性によってストーリーが変わる映画
(ロマンスが人気の属性にはロマンス展開多め等、他ジャンルとの組合せアニメ＋ロマンス＋コメディなど)

実は！アラフィフにアニメが人気！？