購買履歴データの特徴量による国籍推定

明治大学菊池研究室4年

田中司

はじめに

背景：個人情報保護法の改正に匿名加工情報が定義された
　　しかし、加工手法が定まっていない

目的：匿名加工手法を採用する上で重要視される
　　匿名化すべき項目を明らかにする

購買履歴データの概要

PWSCUP2016匿名加工・再識別コンテストのデータセット

顧客データの例

顧客ID	性別	生年月日	国籍
1	f	1995/9/6	Japan
2	m	2000/6/2	China

購買履歴データの例

顧客ID	伝票ID	購入日付	購入時間	商品ID	単価	購入数
1	A	1995/9/6	6:00	12	1.2	5
2	B	2000/6/2	12:00	13	2.4	6

総合データの生成

購買履歴データを用いて国籍ごとに分析するが国籍が含まれていない

顧客データに含まれる顧客IDの国籍を購買履歴データに結合させる

これを本研究では総合データと定義と呼ぶ

総合データの例

伝票ID	購入日付	購入時間	伝票ID	単価	購入数	性別	生年月日	国籍
A	1995/9/6	6:00	12	1.2	5	f	1995/9/6	Japan
B	2000/9/3	12:00	13	2.4	6	m	2000/6/2	China

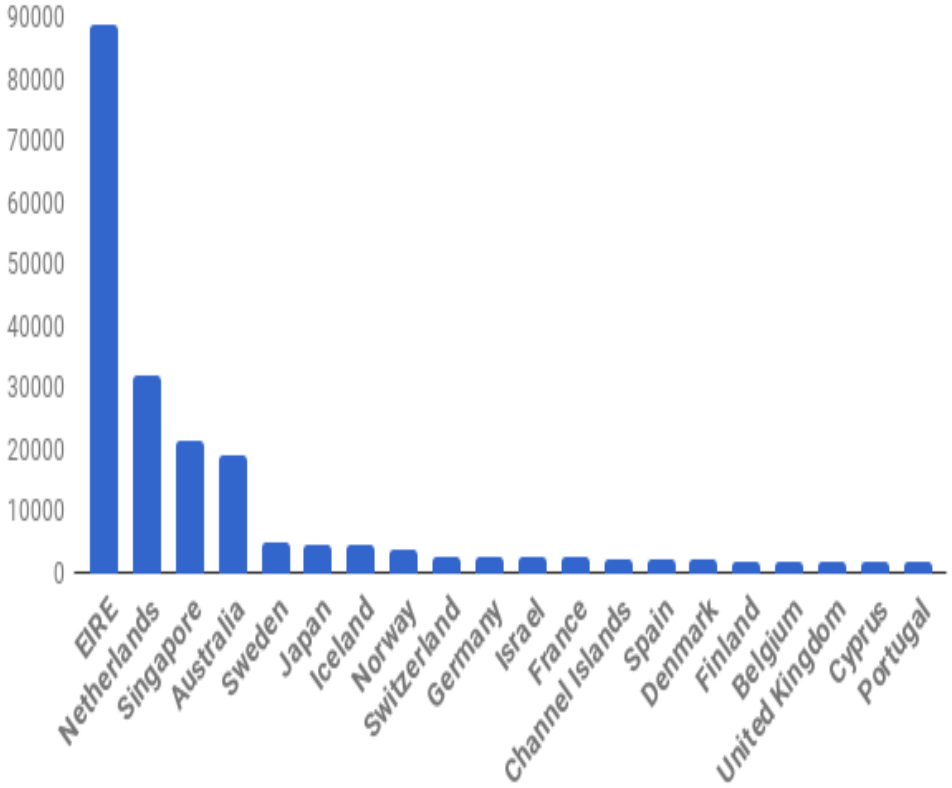
国毎の分析1

総合データを用いて国毎に分析する

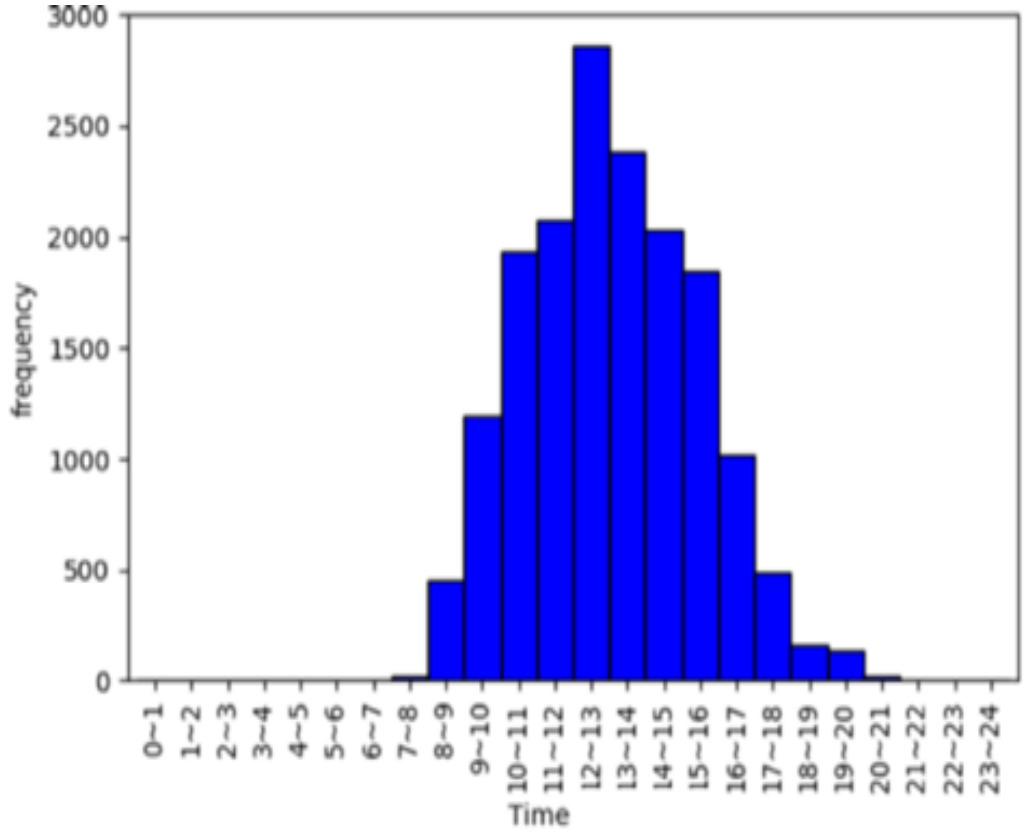
各国の顧客数,合計購入額,取引商品数,伝票数の例

国名	顧客数	合計購入額	取引商品数	伝票数
United Kingdom	3918	7231007.31	3643	16634
France	87	209024.05	1522	389
Belgium	24	44389.61	750	104
Netherlands	9	285446.34	782	94
Germany	94	228867.14	1664	457

国毎の分析2



顧客あたりの合計購入額トップ20国



購入回数の時間帯分布(United Kingdom)

分析結果

- 顧客あたりの購入額トップ20国

最も高いのがアイルランドであるが顧客数が3人

UnitedKingdomの顧客数3918で**特異な顧客による影響**の可能性がある

- 購入回数の時間帯分布

購入頻度はどの国においても9時から18時の間に集中

国籍の時差は見られなかった

国籍が異なる顧客でも在住している国はヨーロッパであるから国の時差が生じないと予想できる

分析による特徴量

月別売上個数について国籍の特徴量から顧客IDの国籍を推定するため以下のように定義する

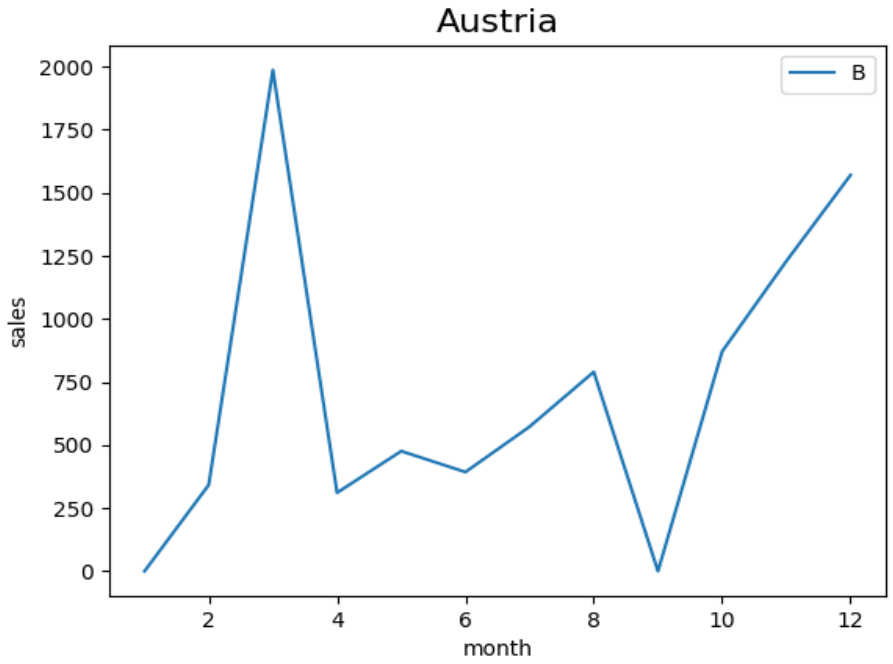
各月の売上が0であるかどうかに着目

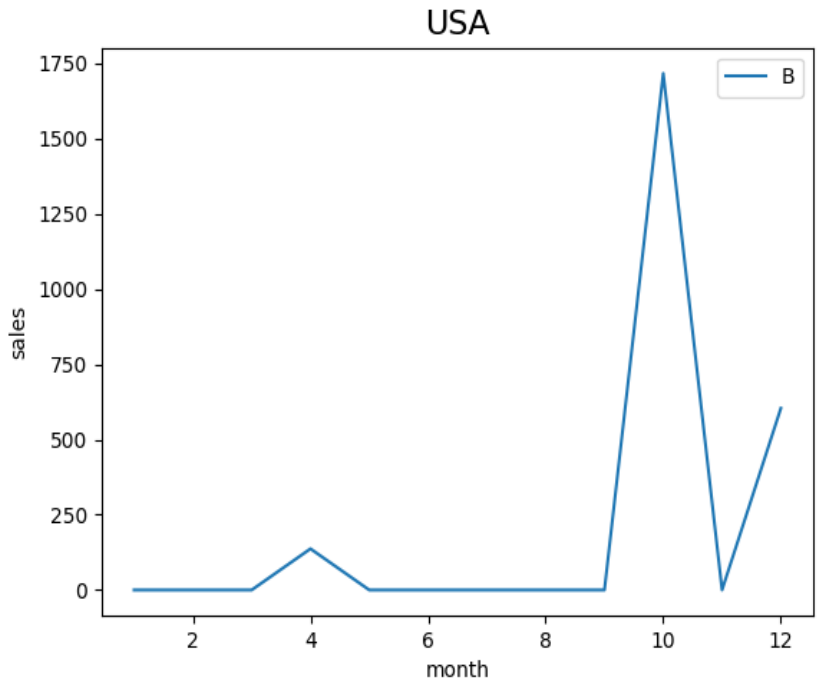
- 月の売り上げ個数が0の場合は特徴がありとして1,その他は0とする
- 2進数12桁で表す
- 10進数に変換した値を国籍の購入特徴ID

購入特徴IDの作成

購入特徴IDの作成例

国籍	10進数	2進数
Austria	2056	100000001000
USA	3834	111011111010





結果：国籍が36種類あり,購入特徴IDは24種類が一意であった。

国籍推定の精度

PWSCUP2016の本戦で提出された加工済み購買履歴データを用いる

国籍の国籍推定IDと一致した顧客IDは90種類あった

精度：全て国籍推定されず、**精度0%**であった

国籍推定手法の評価

結果：国籍の購入履歴の特徴が加工済み購買履歴データの顧客IDの特徴と一致する確率は極めて低い

予想される理由：国籍は顧客の集合の売上個数であるので必ずしも国の特徴とその国籍の顧客の特性は一致しない

【例】Austria 100000001000(2)=2056(10)⇦購入特徴ID

オーストリア国籍の顧客ID:606042076の場合

101111111111(2)=4093(10)となり

購入特徴IDがサウジアラビアとなり推定失敗

おわりに

研究結果

- 加工購買履歴データから顧客IDの国籍を推定する事は極めて難しい
- 本研究で用いたデータは国による顧客数に偏りがあり,全ての国に対してその国独自の特徴であるといえる顧客人数ではない

今後の課題

国籍の候補を絞り出し、他の複数の国籍の特徴を組み合わせて精度を向上することが今後の課題