購買履歴データの特徴量による国籍推定

田中司†
明治大学総合数理学部 先端メディアサイエンス学科 菊池研究室†

1. はじめに

個人情報保護法は 2005 年 4 月に施行されて以来すでに 10 年以上経ち，急速な情報通信技術の発展から当初では予想できなかった問題が顕在化してきた．それに伴って 2015 年 9 月 3 日に個人情報保護法が改正された．そこで新設された制度は，特定の個人を識別することができないように個人情報を加工したものを匿名加工情報と定義し，その加工方法を定め，個人情報を第三者に提供する際に匿名加工を施せば個人の許可を必要としない．これは情報の利活用を容易にするものである．しかし，扱うデータによって再識別されやすい項目は異なり，加工手法の種類が定まっていない．匿名加工手法を採用する上で重要視する匿名化すべき項目を明らかにする必要がある．そこで，本研究では扱うデータを具体的に定め，そのデータの各項目を分析し，特徴量を得る．その特徴量によって属性推定手法を提案し，精度を確かめる．

2. 購買履歴データの分析

2.1. 購買履歴データの概要

本研究は， PWSCUP2016 匿名加工・再識別コンテストで用いられたデータセットを扱う．データセットには表 1 の顧客データと表 2 の購買履歴データの 2 種類ある．

2.2. 国毎の分析方法

国毎に分析するにあたり，表 2 の購買履歴データには国籍の項目がない．よって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て分析する．そこで購買履歴データと顧客データに対して，顧客 ID を結合時のキーとして結合させる．結合させた後のデータを総合データと呼び，その例を表 3 に示す．総合データを用いて国毎の月別売上個数，時間帯に対する購入頻度，顧客数，合計購入額，顧客あたりの購入額，商品種類数，伝票種類数，顧客あたりの平均伝票数，合計個数，伝票あたりの平均個数，レコード数，平均購入額，平均購入月を求める．

2.3. 分析結果

顧客あたりの購入額トップ 20 国を図 1 に示す．顧客あたりの平均購入額の最も高いのがアイルランドであるが顧客数は 3 人で United Kingdom の 3918 人と比べると少ない．よってこの分析結果からアイルランドは顧客合計購入額が多いとは言えず，特異な顧客による影響である可能性がある．また時間帯による購入頻度についてはどの国においても 9 時から 18 時の間に集中していて，国の時差によって生じる購入する時間帯の差は見られなかった．おそらく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．United Kingdom の購入回数の時刻分布を図 2 に示す．各国の顧客数，合計購入額，取引商品数，伝票数を表 4 に示した．

表1 顧客データ

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

表2 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

表3 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

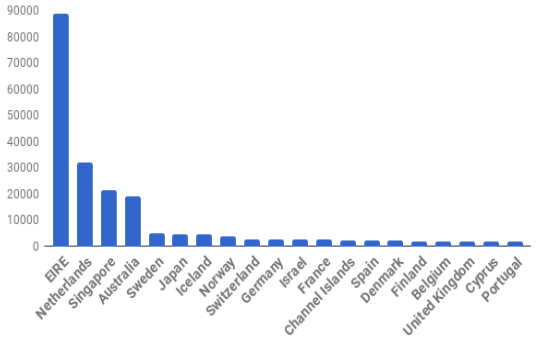

図1 顧客あたりの合計購入額トップ20国

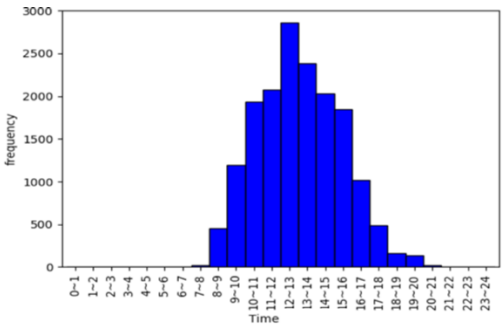

図2 購入回数の時間帯分布(United Kingdom)

表4 使用したデータの概要

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
Norway	10	36165.44	591	36
EIRE	3	265545.9	1943	260
Switzerland	21	57222.85	951	55
Spain	29	64021.95	1139	94
Poland	6	7334.65	204	19
Portugal	19	33439.89	686	57
Italy	14	17483.24	473	38
Lithuania	1	1661.06	29	4
Japan	8	37416.37	215	19
Iceland	1	4310	103	7
Australia	7	132620.44	543	42
Channel Islands	9	20450.44	430	26
Austria	11	16779.59	472	23
Sweden	8	38378.33	261	36
Finland	12	22546.08	458	41
Cyprus	6	10697.15	418	15
Greece	4	4760.52	138	5
Singapore	1	21279.29	178	7
Lebanon	1	1693.88	45	1
United Arab Emirates	2	1902.28	68	3
Israel	3	7221.69	219	5
Denmark	7	14751.52	204	14
Saudi Arabia	1	145.92	9	1
Czech Republic	1	826.74	25	2
Canada	4	3666.38	147	6
Brazil	1	1143.6	32	1
USA	4	3580.39	163	5
European Community	1	1300.25	50	4
Bahrain	2	548.4	16	2
Malta	2	2725.59	99	5
RSA	1	1002.31	57	1

表5 国籍と購入特徴IDの例

国籍	10 進数	2 進数
Denmark	9	000000001001
Norway	25	000000011001
Austria	257	000100000001
Poland	512	001000000000
USA	1527	010111110111
Cyprus	2076	100000011100

3. 属性推定

3.1. 分析による特徴量

月別売上個数についての国の特徴量を顧客の購買履歴から国籍の推定を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する。その月の売り上げ個数が 0 の場合は特徴がありとして 1,それ以外は 0 とする.それらを 2 進数 12 桁で表し,10 進数に変換した値を国による購入特徴 ID と定義する.表 5 に国と購入特徴 ID の例を示す.例えば USA の場合,売上個数が 0 である月は 1,2,3,5,6,7,8,9,11 の月であり,2 進数で 111011111010,10 進数に変換すると 1527 であり,これが USA の購入特徴 ID である.

3.2. 国籍推定手法

3.1節で定義した再識別IDを利用してそれぞれの国籍に対して購入特徴IDをプログラム言語Pythonでプログラムを作って国籍推定を試みる.国籍数は36種類あり,再識別IDが一意である国が24種類あった.顧客IDに対しても月別売上個数から再識別IDを割り出し,国籍と顧客IDの再識別IDが一致すればその国籍であると予想できる.表5に国籍と購入特徴IDの例を示す.

3.3. 国籍推定の精度

実際のPWSCUP2016の本戦で提出された加工済み購買履データを用いて国籍再識別手法の精度を確かめる.まず加工済み購買履歴データのすべての顧客IDの月別売上個数を求め,再識別IDを算出する.国籍の再識別IDと一致した顧客IDは90種類あった.実際の国籍と比較すると90種類全て再識別されなかった.精度は0%である.ある国の再識別ID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確率であることがわかった.

4. おわりに

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かった.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.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かったことが原因と考える.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である.

参考文献

[1] 経済産業省　商務情報政策局　情報経済課,“個人情報の保護に関する法律についてのガイドライン”（http://www.meti.go.jp/policy/it_policy/privacy/ 2016 12 月参照）

[2] 菊池浩明, 小栗 秀暢, 野島 良, 濱田 浩気, 村上 隆夫, 山岡 裕司, 山口 高康, 渡辺 知恵美,“PWSCUP:履歴データを安全に加工せよ”, CSS2016, pp.271-278, 2016.