明治大学総合数理学部

2016 年度

卒　業　研　究

購入履歴データを用いた匿名加工データの最頻アイテムに注目した再識別手法 freqItem の提案とその評価

学位請求者　先端メディアサイエンス学科

岡本健太郎

目次

1 はじめに 4

1.1 研究目的 4

1.2 手法のモックアップ 4

1.3 目的の解決の手段 4

2 匿名加工・再識別 5

3 プチ PWSCUP について 6

3.1 禁止事項 6

3.2 使用データ 6

3.2.1 定期範囲 6

3.2.2 回数 6

3.2.3 乗降駅 7

3.2.4 乗降路線 7

3.2.5 使用用途 7

3.2.6 使用場所 7

3.3 施した加工 8

3.4 結果 8

4 用いたデータについて 9

4.1 元データとコンテストデータ 9

4.2 PWSCUP について 9

5 データの調査について 10

5.1 Suica データの分析 10

5.2 購入履歴データの分析 11

6 再識別手法 freqItem 13

6.1 コンテストで使われた加工手法について 13

6.1.1 YA-匿名化 13

6.1.2 Jaccard 最適化 13

6.1.3 Jaccard ランダム化 13

6.1.4 最適化 13
6.1.5 仮名化 14
6.1.6 列統一 14
6.1.7 属性値追加 14
6.1.8 グループ内スワップ 14
6.2 再識別手法 freqItem について 15
6.3 再識別手法 Jaccard 再識別について 16
7 再識別結果 17
8 考察 18
9 おわりに 19
謝辞 20
参考文献 21
付録　ベイジアンネットを用いたサッカーの有効戦術推定 22
10 概要 23
11 はじめに 24
12 提案手法 25
12.1 ベイジアンネット 25
12.2 ノードの決定・定義 25
12.3 DAG の構築 27
12.4 学習結果 28
13 おわりに 31
参考文献 32

第１章　はじめに

近年，ビッグデータをデータマイニングする取り組みが盛んになってきているが，データの一部を削除しても個人が特定されてしまうような事態が不安視されている．そこで，2015 年に匿名加工に関する個人情報保護法が改正されこのようなプライバシー保護の観点から匿名加工という概念が生まれた．

本稿では，情報処理学会の CSS(Computer Security Symposium)にて開催された匿名加工コンテスト，通称 PWSCUP2016 において各チームによってさまざまな加工をされた購買データを用いて本提案手法である最頻アイテム再識別にて再識別を行い，提案した再識別率を評価し，他の手法との比較を示す．

1.1　研究目的

偽造タプルの挿入，またはタプルの削除を最小限に抑えた動的データの匿名加工の手法を実装する．

1.2　提案方式のモックアップ

表１　従来手法　　表２　提案手法

職業	病気
国語教師	風邪
弁護士	風邪
数学教師	骨折
検事	骨折
理科教師	風邪
裁判官	風邪
社会教師	骨折
検事	骨折
弁護士	頭痛
弁護士	頭痛

⇒

職業	病気
教師	風邪
法律関係	風邪
教師	骨折
法律関係	骨折
教師	風邪
法律関係	風邪
教師	骨折
法律関係	骨折
法律関係	頭痛
法律関係	頭痛

職業	病気
国語教師	風邪
弁護士	風邪
数学教師	骨折
検事	骨折
理科教師	風邪
裁判官	風邪
社会教師	骨折
検事	骨折
弁護士	頭痛
弁護士	頭痛

⇒

職業	病気
教師	風邪
法律関係	風邪
教師	骨折
法律関係	骨折
教師	風邪
法律関係	風邪
教師	骨折
法律関係	骨折
法律関係	頭痛
法律関係	頭痛
教師	頭痛
教師	頭痛

従来手法[3]では m-不変性を満たすように加工すると教師グループが 2-不変性，法律関係グループは 3-不変性であるが，提案手法は教師グループが 3-不変性，法律関係が 3-不変性を満たすように偽造タプルを挿入した．以後，このようにすべてのグループに対して p-不変性を満たすような性質を p-統一性と呼び，p-統一性を持つように動的データを加工することによってデータに含まれるあらゆるグループに対して再識別される可能性を 1/p に統一し，再識別のリスクを最小限に抑える．

1.3　解決の手段

動的データのシグネチャとその個数の集合を{S}={(a,5),(b,3),(c,2),(d,2),(e,1)}のように定め，{S}に含まれるシグネチャとその個数の集合を{S}*={1,2,2,3,5}で定める．{S}*の最小値(1)，最大値(5)，中間値(2)を割り出し，中間値 n に対して n-統一性を持たせるようなデータの加工をする．例えば今回の例では中間値は 2 であるので 2-統一性を持たせるようにシグネチャ e は偽造タプルの挿入，a，b に対してはどのグループにも属さないタプルに関して削除を行う．これにより偽造タプルの挿入，削除を最小限にしてデータを加工することが可能である．

第 2 章　匿名加工・再識別

本研究において，再識別の手法を述べる際にまず匿名加工の基礎知識を述べることとする．匿名加工とは、ビッグデータに含まれる個人のプライバシーを守りつつデータの有用性を保つようなデータ加工である．また，ビッグデータは名前やマイナンバーのような個人を直接的に特定しうる属性を ID，誕生日や住所，性別など組み合わせることで間接的に個人を特定しうる属性を QI，その他の保護すべき対象の属性を SA と定義し，各属性をこの 3 要素に分類する．匿名加工において ID は当然削除対象なのだが，QI を組み合わせることによって個人を特定できる場面がしばしば生じる．

そこで k-匿名化やトップコーディングなどといった加工を施すことにより QI からも個人を特定できないようにするのが主流である．また，その加工データは有用性と安全性という二つの

観点から評価をされ，その価値を問われる．加工データの有用性を示す指標には MAE，安全性を示す指標には l-多様性や m-不変性などがある．これらの指標は同じ QI を持つグループに分類したとき，どのグループの平均値も元データと近似するかどうかや，どのグループの SA が l 個，または m 個持つように加工を施されているかを示す指標である．

再識別とは匿名加工されたデータから個人を特定するための一連の手順のことを指す．特に，K-匿名化された加工データにおいては個人を再識別できる確率は 1/K になる．一般的に匿名加工されたデータにおいて L-多様性や M-不変性を満たすためにノイズを加えることがあるが，加工を行うことで安全性が増す代わりに有用性が下がることが知られており，今日の匿名加工の分野では有用性を下げないように安全性を高める手法の提案が議論の中心である．

第 3 章　プチ PWSCUP について

プチ PWSCUP とは，菊池研究室内で匿名加工を研究している班員同士で匿名加工技術の向上を目的とした試みであり，CSS で行われている PWSCUP（Privacy Work Shop CUP）を参考にして菊池教授の指導の下で行われた．データは菊池研究室に所属する教員，学生 34 人分の Suica の過

去 20 件分の利用履歴データを用いた．また，期間は 8 月 6 日から２６日までに班員が 10 件以上加工データを提出し，有用性指標と再識別率から順位を決めた．

3.1　禁止事項

他班員との結託，他班員へのなりすまし，指標の内容の偽装，不正な指標の実装，他班員のデータ・指標の削除，データのフォーマットや属性の型の変更を禁止事項とした．

3.2　使用データ

菊池研に所属する教員 1 名，学生 31 名合計 32 人分の Suica または PASMO についてカードリーダーを用いて利用履歴データを作成した．そのデータの詳細は表 3 に記した．また，名前や性別はそのままの意味であるが，私たちが独自に定めた属性の定義を 4.1.1 以降に記す．また，データ収集にはカードリーダーを用いたのだが，データのスキーマについては未調査であるため，本節には定期券を用いた実験に基づいた推測が含まれていることをここに明記する．

表 3　Suica の利用履歴データの詳細

	マスターデータ	トランザクションデータ
含有情報	Suica 所有者の情報	Sucia の利用履歴を日付ごとに追加したデータ
行数	34 行	585 行
属性	名前，性別，学年，住所，定期範囲 1・2 の 6 属性	行番号，名前，日付，回数，乗車駅，降車駅，乗車路線，降車路線，使用用途，使用場所，料金の１２属性

3.2.1　定期範囲

定期範囲１は自宅の最寄り駅，定期範囲２は学校の最寄り駅を設定した．つまり，このデータでは間接的に定期範囲２は中野を指している（今後データ収集の規模を大きくする際にはこの限りではない）．Suica の乗降履歴には定期範囲が印字されないので履歴データだけでは正確な移動範囲が分からないため設定した属性である．

3.2.2　回数

ユーザーのその日の Suica の使用が何回目かを示す指標である．ただし，Suica の利用は改札を通すだけでなく，コンビニや自動販売機での物販購入や料金のチャージも含まれている．

3.2.3　乗降車駅

改札を通した駅を指している．また，この属性には駅の名前ではなく駅それぞれに振られた id が格納されている．例えば，渋谷駅から新宿経由で中野駅まで行ったとき，乗車駅は渋谷，降車駅は中野になるのだが定期範囲が渋谷から新宿だった場合，乗車駅は変わらず渋谷，降車駅は中野から新宿になる．また，3.2.5 節における定義で使用用途が交通でないときにはこの属性には値が入らない．

3.2.4　乗降車路線

改札を通した路線を指している．例えば，3.2.3 同様渋谷から中野まで行った場合，渋谷から新宿までは JR 山手線，埼京線，湘南新宿ラインの可能性があるのだが，JR 山手線が選択される．この理由についてはおそらくどの電車に乗ったかはさほど重要でなくどの会社の路線を選択しているかであるので，最も一般的な路線が選択されていると考えられる（本データには JR 埼京線，湘南新宿ラインは含まれていなかった）．また，定期範囲が 4.2.3 同様で新宿から秋葉原へ行った場合，中央線に乗っていたにも関わらず乗車駅は代々木，降車駅は秋葉原，路線は総武線が選択されていた．これはまず乗車範囲から定期範囲を削除したのちに矛盾しないような路線が選択されているからであると考えられる．また，乗降車駅同様，使用用途が交通でないときこの属性には値が入らない．

3.2.5　使用用途

本データには主に交通，物販，チャージ，バスチャージ，共通という 5 つの使用用途が見られたため離散値で表現した属性である．交通は電車の利用，物販は Suica を用いて自動販売機などを利用したとき，チャージは切符の券売機によるチャージ，バスチャージはバスでのチャージ，共通は主にバス利用による料金を指している．また，交通，共通以外の項目に関してはその使用

目的を 3.2.6 節にて述べている.

3.2.6　使用場所

主に交通機関の利用以外に Suica を用いた場合にこの属性に値が格納されている. 目的には 8 項目あり, 自販機・物販端末・精算機・車載端末・券売機等・券売機・乗り継ぎ精算機・簡易金機であった. 券売機等と券売機の違いについては未調査であるがおそらく前者は飲食店などの券売機, 後者は切符の券売機であるとユーザーの使用感から推測された. 調査済みのものに関しては, 物販端末はコンビニなどの支払い, 車載端末はバスなどについているもの, 簡易金機は郊外の無人駅などに設置されているチャージをするだけの機械を指している.

3.3　施した匿名加工

私が施した加工を表 4 に示した. また, 加工は excel の関数を用いて行ったため属性全体に一様の加工を施す結果となった.

表 4　提出したデータの加工方法と加工による保護対象

データ名	加工方法	保護対象
Copy	降車駅を乗車駅にコピー	定期範囲
plusRandNum	料金に[-9,9]のノイズを付加	料金の合計
deleteDigit	料金の 3 桁目以降を削除し, 2 桁に丸める	料金の合計
changeUse	用途が物販の列をユーザーごとに最も利用した駅同士の乗降履歴に変換する	入出場回数
unityUse	使用用途の列を物販に統一	用途
Booleanize	日付を休日と平日に二値化	日付グループの乗降数
Plus7	日付を 7 日後にずらす	日付グループの乗降数
unityIn	乗車駅を新宿に統一	定期範囲
randUser	最も少ないユーザーのすべてのタプルをノイズ化	ユーザーグループ

3.4　結果

結果は残念ながら最も順位が高いデータが unityIn の 9 位という結果になった．順位の高いデータは，似ているユーザーを抽出し，グループを作ってそのグループごとに加工を施していた．また，この結果を受けて有効な加工方法を提案するためには闇雲に加工を施すのではなく再識別によるリスクを考えたうえで行うべきだと考えたため，この段階で方向性を匿名加工の研究から再識別の研究へとシフトした．

第４章　コンテストで用いたデータについて

匿名加工には静的データと動的データの 2 種類が存在する．含有されるタプルが時間によって変化しないものを静的データ，変化するものを動的データとする．例えば，本大会 PWSCUP2016 で扱ったオンラインショッピングの履歴データは動的データである．

4.1　元データとコンテストデータ

今回 PWSCUP2016 で用いたデータは，UCI データセットのオンラインショッピングサイトの購買データにおいて顧客数を 10%にサンプリングしたデータを用いている．元データを構成する要素を表 5 に示す．また，顧客の個人情報において性別や生年月日は乱数を用いて合成されている．

このデータセットは顧客個人の情報が入ったマスターデータ，購入履歴が入ったトランザクションデータで構成され，元マスターデータ 4333 顧客中 400 人を無作為に抽出したものをコンテストマスターデータ，その 400 人が含まれる行だけを残したものをコンテストトランザクションデータと呼ぶ．

表 5　データの詳細

	マスターデータ	トランザクションデータ
含有情報	顧客の情報	各顧客の購買履歴
属性	顧客 ID, 性別, 生年月日, 国籍の 4 属性	顧客 ID, 伝票 ID, 購買日時, 購買時間, 製品 ID, 単価, 購買数の 7 属性
行数	4333 行	397625 行
コンテストデータ		
行数	400 行	38087 行

4.2　PWSCUP について

匿名加工されたデータは有用性と安全性に二つの観点からその価値を評価され, PWSCUP2016 では有用性指標を平均絶対誤差 CMAE, ハミング距離 ham, 購買アイテム集合 topitem の 5 つ, 安全性指標を再識別されたユーザー数の割合と定義した[5]. コンテストは予備戦と本戦に分かれていてそれぞれでデータを加工する匿名加工フェイズ, チーム同士でそれぞれ提出されたデータを再識別する再識別フェイズを行い, 予備戦本戦の結果が 1：9 の割合で順位を決定する. 匿名加工フェイズでは顧客マスターデータ M と履歴トランザクションデータ T が配布される. そして各々の手法で加工し, 加工マスターデータ M′ と加工トランザクションデータ T′ , そしてマスターデータの行番号データ P を提出し, 各有用性指標の最大値をそのデータの有用性の値とした. その後再識別フェイズではコンテストサイトにおいて各チームの M′ と T′ が公開されているので各自ダウンロードし再識別を行ったのち, 行番号データ Q のみを提出する. そこで P と Q を照合し, 正解率を再識別率の値とした.

第５章　データの調査について

本節では, データの有効な加工を審査するにあたってプチ PWSCUP, PWSCUP2016 で使われたデータについて調査をした結果を報告する.

5.1　Suica データの分析

Suica データは最頻乗降駅, 最頻利用日, ユーザーごとの平均利用額について調査した. 図１・

２・３にそれぞれ最頻乗降駅，最頻利用日，ユーザーごとの平均利用額のグラフを示した．最頻乗降駅は新宿 89 回，中野 75 回，渋谷 47 回，高田馬場 45 回という結果になった．また，最頻利用日は 2016/5/28 であり，図 2 を見てみると利用頻度に周期があることが分かった．ユーザーごとの平均利用額は最も菊池教授が多かった，また就職活動をしている人は平均利用額が高くなる傾向も見られた．例外として伊藤氏は定期を所持していないため利用額が多い結果となった．

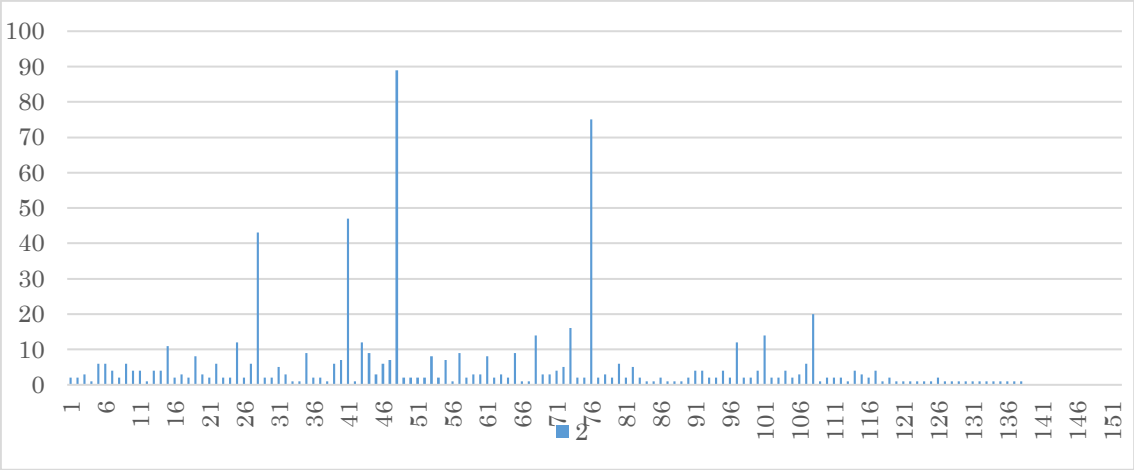

図１　駅の利用回数グラフ

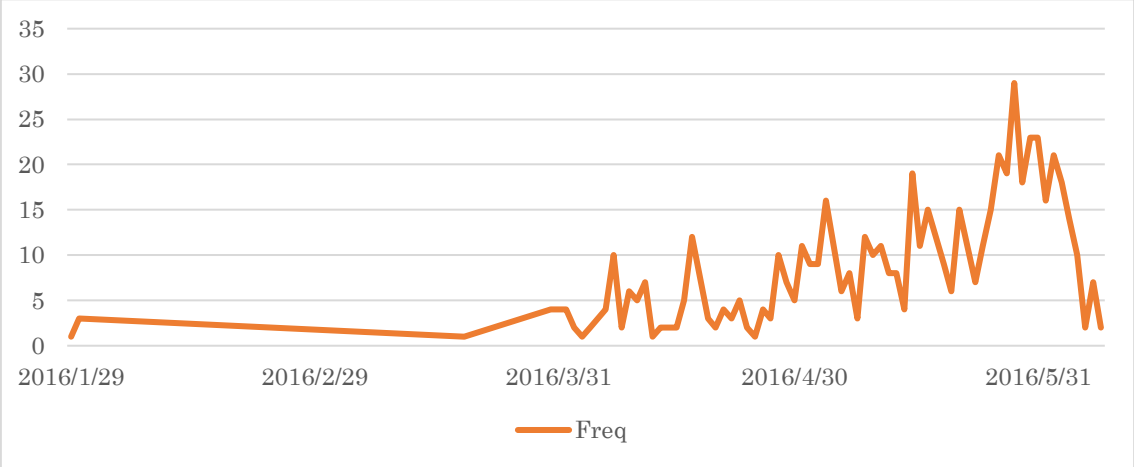

図２　日にちごとの利用頻度グラフ

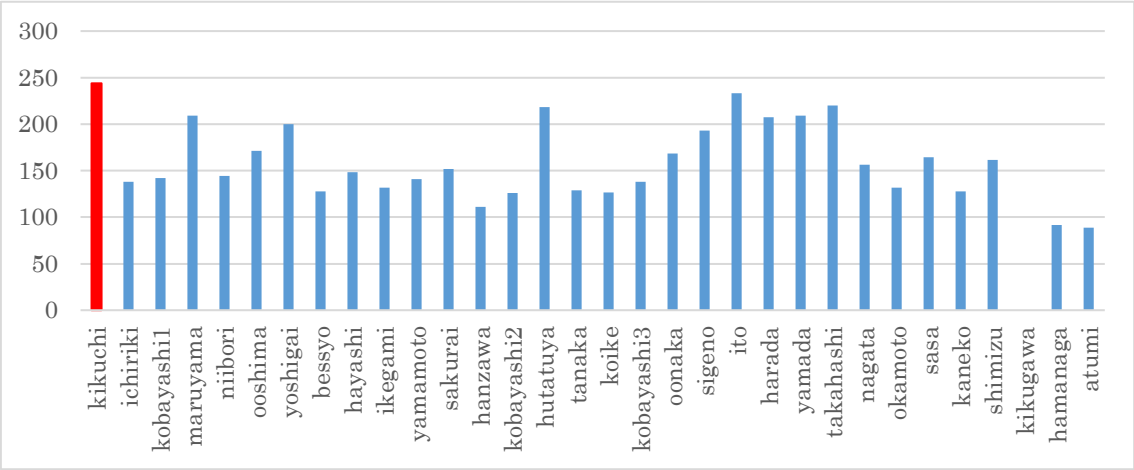

図３　ユーザーごとの平均使用額グラフ

5.2　購入履歴データの分析

購入履歴データは顧客ごとの購入頻度と購入額の比較，利用日ごとの購入回数と購入総額の比較を行った．図４・５にそれぞれ顧客ごとの購入回数と購入総額の比較，利用日ごとの購入回数と購入総額の比較のグラフを示した．それぞれのグラフではオレンジの棒グラフが購入頻度，青の棒グラフが購入総額を示している．どちらのグラフもオレンジと青のグラフがほぼ重なっているのである程度の相関関係があることが分かった．図４の特徴的なデータとしては利用回数の最も多いユーザーと購入総額が最も多いユーザーがそれぞれ違うことと，一度だけ大量額を購入したユーザーがいたことである．図 5 の特徴的なデータとしてはグラフが跳ね上がるのは月の中旬であることと 2011/1/18 にだけ爆発的な購入があったことである．

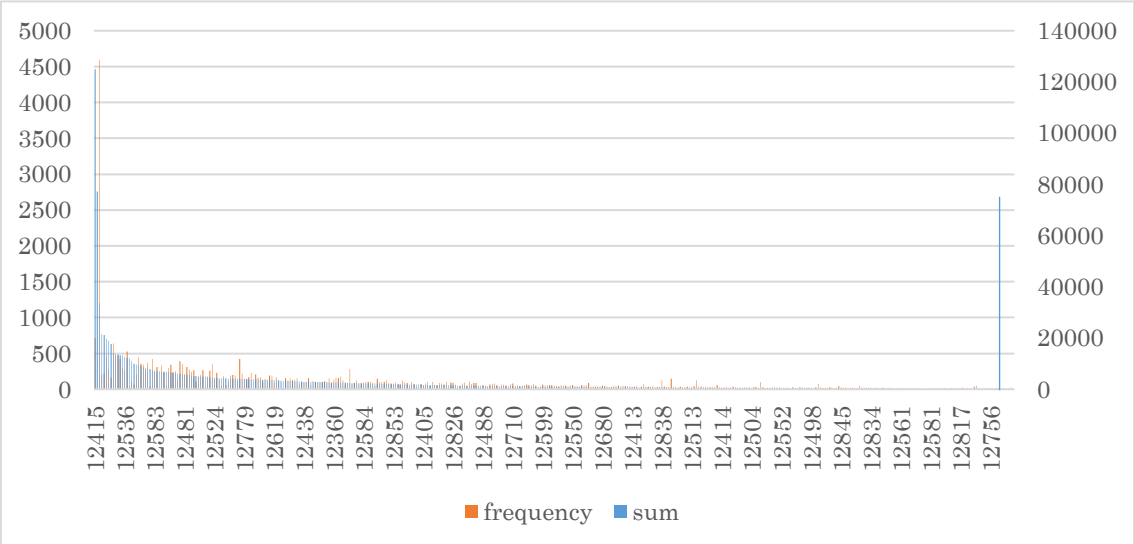

図４　顧客ごとの購入回数と購入総額の比較

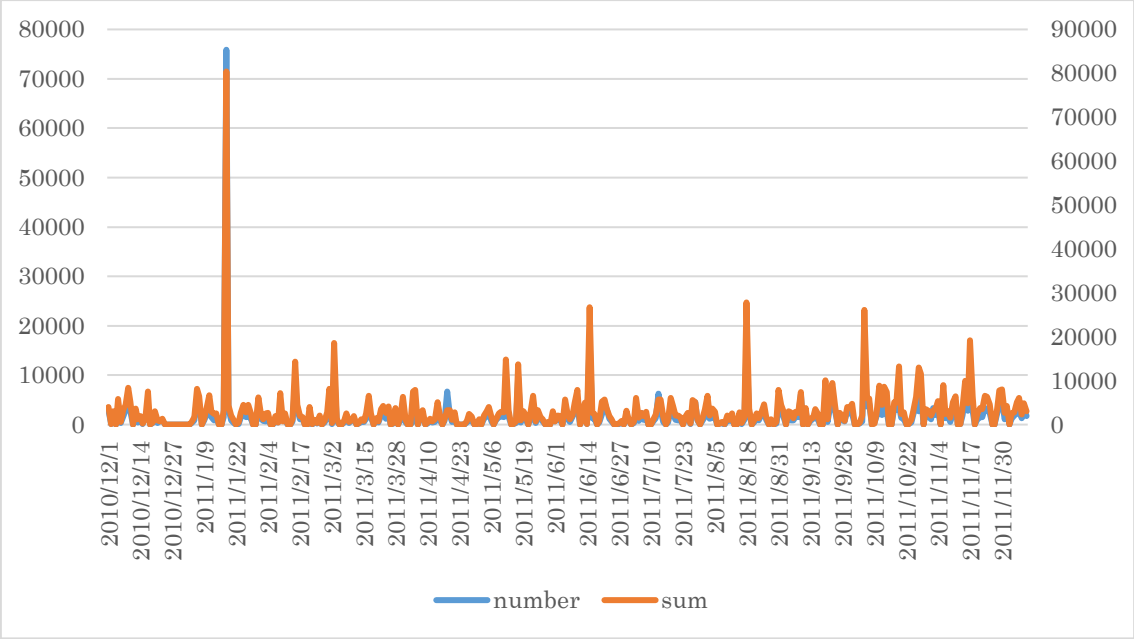

図５　利用日ごとの購入回数と購入総額

第 6 章　提案手法 freqItem

本研究では，有用性指標の中でもとりわけ Y2-jaccard を下げないような匿名加工データを対象にした再識別手法 freqItem を提案する．Y2-jaccard とは購入データの中でユーザーごとに購入アイテムで多重集合を形成し，この集合の差が加工前と加工後で大きいと過加工とみなされ排除されてしまう．この指標は YA-匿名化という，PWSCUP において有用性を全く下げることなく安全性を高める手法を防ぐために用意された指標である，したがって，各チームが最も気にした指標がこの Y2-jaccard であり，かつ商品集合は加工をしづらいファクターであるため最も高い効果が得られると私は考えた．

6.1　コンテストで使われた加工手法について

本コンテストでは，事前に定められた有用性指標スクリプトに加工データを入力として値が出力されることによってそのデータの価値を定めている．そのため，闇雲にノイズを加えるのではなく有用性指標を下げないような加工法を提案することが重要である．そこで，本節ではどの有用性指標を対策しているかに注目しながら用いられた主な手法を紹介する．

6.1.1　YA-匿名化

YA-匿名化とは M＝M′，T＝T′ となるようなデータを提出するのだが，行番号データ P をランダムに入れ替える加工手法である．元データとまったく同じデータのため，どの有用性指標も変化せず，その有用性指標は最小値 0 を示し，すべてのユーザーを特定できたとしても P がランダムに入れ替わっているため再識別率は大幅に下がってしまう．

6.1.2　Jaccard 最適化

Y2-Jaccard は 6.1.1 節で記した YA-匿名化を防ぐための足切り指標であるが，YA-匿名化をしながら Y2-Jaccard が足切りにかからないようなギリギリの YA-匿名化データを探索するのが Jaccard 最適化である．

6.1.3　Jaccard ランダム化

4.1.2 節では Y2-Jaccard が足切りにかからないギリギリの探索を行う手法であったが，最適化を行うとそれが手掛かりとなってしまい，YA-匿名をしたにも関わらず再識別されてしまう恐れがある．それを防ぐのが Jaccard ランダム化であり，最適化はしないが最適化された Y2-Jaccard の値に近いデータを選択することで Jaccard 最適化されたデータよりも再識別されにくい YA-匿名化をする．

6.1.4　最適化

CMAE1，CMAE2，ut-rfm などの有用性指標を最小にするようなトランザクションデータの加工を探索する加工手法であり，YA-匿名化だけでなく偽造タプルの挿入やノイズを加えた後にこれらの有用性を回復するようなタプルの挿入などが最適化に当てはまる．

6.1.5　仮名化

トランザクションデータには ID を示す列があるため，ID の列を見ればどのユーザーをどの程度加工したのかが分かってしまう．それを防ぐのが仮名化であり，マスターデータの顧客 ID に対してハッシュ関数を用いて仮 ID を割り振り，トランザクションデータの ID の列を仮 ID で表現することで元データの顧客と結び付けられないようにする加工法である．表 6，表 7 は仮名化する前と後のマスターデータを表したものである．この例を元にすると Abe→b，Baba→d，Chiaki→c，Doi→a のように ID が変換されているがタプルの内容は変化していないことが分かる．

表 6　仮名化前のマスターデータ

ID	Country	Sex	Birthday
Abe	America	M	Feb
Baba	Bulgaria	M	Dec
Chiaki	Canada	F	May
Doi	Denmark	M	Aug

表 7　仮名化後のマスターデータ

ID	Country	Sex	Birthday
B	America	M	Feb
D	Bulgaria	M	Dec
C	Canada	F	May
A	Denmark	M	Aug

6.1.6　列統一

time の列に関する有用性指標が定められてないためどれほどの加工を施してもどの有用性指標も下がらない．したがって time の列の値をランダムな値に統一する加工である．ランダム化でも結果はほとんど変わらない．

6.1.7　属性値追加

元データの集合に含まれていなかった属性値を追加する手法である．CMAE や RFM の有用性は元データの集合に対して平均絶対誤差などをとるため,新たな値に関しては平均絶対誤差は 0 となる．

6.1.8　グループ内スワップ

グループ内スワップは有用性指標を下げないようにグループ内で値だけをユーザー同士で入れ替える手法である．例えば，表 8 は表 6 のユーザーを性別グループ内で country の列をスワップした例である．例えば，元データでは Abe の country は America であるが表 8 でスワップした後は Bulgaria になっている．なお，Birthday の列の値は変化していない．

表 8　性別グループスワップの例

ID	Country	Sex	Birthday
Abe	Bulgaria	M	Feb
Baba	Denmark	M	Dec
Chiaki	Canada	F	May
Doi	America	M	Aug

6.2　再識別手法 freqItem について

freqItem とは最頻アイテムに注目した再識別手法である．Y2-jaccard は足切指標であるので加工しづらい要素である．しかし，上位チームは Y2-jaccard に関して当然有効な加工をしてくるであろうと考え，Y2-jaccard と同様の freqitem という概念を提唱する．freqitem とはユーザーごとに最も購入したアイテムを単一に定めた値であり，全ユーザーの人気商品の多重集合 topitem とは異なり，最頻アイテムが複数存在した場合はその中でランダムに決定する．例えば，表 9 において，topitem が{doughnut，eraser}だとすると各ユーザーの topitem と freqitem は表 10 のようになる．どのユーザーにおいても，topitem に freqItem に含まれている．topitem とは item 集合の中で頻度の高いアイテムであるのでこのような結果になることが多いと考えられる．例えば，匿名性を高めるために，2 行目を削除すると topitem は eraser が削除され doughnut のみとなるが freqItem は変化しない．

表 9　履歴データの例

	Name	Item
1	Abe	Doughnut
2	Abe	Eraser
3	Baba	Doughnut
4	Abe	Doughnut
5	Chiaki	Fork
6	Baba	Eraser
7	Chiaki	Eraser
8	Baba	Eraser

表 10　ユーザーごとの topitem と freqItem

Name	Topitem	freqItem
Abe	doughnut，eraser	Doughnut
Baba	doughnut,eraser	Eraser
Chiaki	Eraser	eraser または fork

6.3　再識別手法 Jaccard 再識別について

Jaccrad 再識別[6]は，多重集合の類似度 jaccard 係数[5]を用いて jaccard 係数をユーザーごとに求め，最も近いユーザーを同一ユーザーと再識別している．なお，この再識別手法は PWSCUP2016 において再識別賞を受賞した．

第 7 章　再識別結果

表 11 は，PWSCUP2016 の各チームの加工手法と各加工データに対する freqItem 再識別と jaccard 再識別の識別率を示している．ここで，チーム名は，上から順にランキングの上位にいることを表している．表 11 に示された通り，jaccard 再識別は freqItem 再識別の再識別率の平均 2.90 倍の精度があることがわかる．とりわけ，チーム T と K においては 10 倍以上の精度があり，チーム I に対しては 88.50%とほとんどのユーザーの再識別に成功している．すべてのチームで仮名化，列統一が，チーム T 以外のすべてのチームで YA-匿名化を採用している．また，上位 3 チームでは Jaccard 最適化およびランダム化を行っている．それと同様に上位 3 チームにおいて freqItem 再識別の識別率が低く，特にチーム T と K に関してはほとんどのユーザーを識別できていない．

表 11　各チームの加工手法と freqItem 再識別・Jaccard 再識別の再識別率の違い

チーム名	freqItem 再識別	Jaccard 再識別	YA 匿名化	Ja 最適化	Ja ランダム化	CM AE1 最適化	CM AE2 最適化	RF M 最適化	仮名化	列統一	属性値追加	国スワップ	国&性別スワップ	購入スワップ
T	1.25	22.25	○		○	○	○		○	○				
K	0.75	25.50	○	○		○	○		○	○				○
J	9.50	27.50	○		○	○	○	○	○	○				○
B	14.75	30.25	○			○	○		○	○			○	○
N	15.25	27.50	○	○			○		○	○	○	○	○	○
M	13.00	38.50	○	○					○	○		○	○	
I	44.75	88.50	○						○	○				
平均	14.18	27.50												

7.1 考察

表 11 の加工方法により，freqItem 再識別は仮名化，YA-匿名化などといった手法には識別率を左右されず，スワップに関しては購入グループ内でのスワップに最も強い．それに対し，jaccard 係数を変えずに行われたアイテムの加工手法 Jaccard 最適化・Jaccard ランダム化に対して極端に弱く，ほとんど識別できないことが分かった．Jaccard 係数を変えずに行われたアイテムの加工手法は topitem 以外のアイテムについて，表 12，13 のような加工が施されていたために識別率が低い結果となった．表 12，13 は購入アイテム集合を変えないように購入したアイテムをユーザーごとに一つだけ残し，他のアイテムはすべて POST に変換した加工である．したがって T の Item の列のほとんどが POST になったことにより，ほとんどすべてのユーザーの freqItem が POST になったので正しく再識別できなかった．そこで，freqItem 再識別に加えて Jaccard 係数も計算し，識別要素とすることでさらに再識別率を高めることができるのではないかと期待している．

表 12　上位チームの加工前の例

	ID	Item
1	Abe	Eraser
2	Abe	Doughnuts
3	Abe	Doughnuts
4	Baba	Eraser
5	Baba	Folk
6	Abe	Eraser
7	Baba	Eraser
8	Baba	Eraser
9	Abe	Folk
10	Baba	Doughnuts
11	Abe	Doughnuts

表 13　上位チームの加工後例

	ID	Item
1	Abe	POST
2	Abe	Doughnuts
3	Abe	POST
4	Baba	POST
5	Baba	POST
6	Abe	Eraser
7	Baba	Eraser
8	Baba	Folk
9	Abe	Folk
10	Baba	Doughnuts
11	Abe	POST

第 8 章　おわりに

本研究では，freqItem 再識別と jaccard 再識別の比較を行うことで手法の評価，ならびに改善策を講じることができた．この結果をもとに，現手法の改善や新たな再識別手法の提案をすることで PPDP の研究の発展に貢献していきたい所存である．

謝辞

プチ PWSCUP のための Suica の利用履歴データの提供に快諾していただいた菊池研究室の皆様，本コンテストデータとその加工手法を提供・公開していただいた PWSCUP2016 参加者のみなさま，Jaccard 再識別を開示していただいた原田氏，ならびに担当教員でありさまざまな助言をしていただいた菊池浩明教授に感謝いたします.

参考文献

[1] 南和宏,“プライバシー保護データパブリッシング”, CSS, pp.1-9, 2013 年 6 月

[2] 菊池亮, 五十嵐大, 濱田浩気, 千田浩司,“データを逐次公開する際のプライバシー保護”NTT, pp.1, 2015 年 4 月

[3] 上土井陽子, 堀内敦史, 沖田梨絵子, 若林真一,“動的データのプライバシ保護再公開における精確な安全性の評価について”SCIS2016, pp.1-4, 2016 年 1 月

[4] 伊藤聡志, 菊池浩明“ユークリッド距離を用いた再識別手法と PWSCup2015 の匿名加工データを使用した評価”, CSEC2016, pp.1-3, 2016 年 5 月

[5] 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匿名加工せよ”, CSS2016, pp.1-7, 2016 年 9 月

[6] 原田玲央, 伊藤聡志, 菊池浩明,“商品の特徴による再識別リスクとクラスタリングを用いた購買履歴データ匿名加工手法の提案”, SCIS2017, pp.2-5, 2017 年 1 月

付録
ベイジアンネットを用いたサッカーの
有効戦術推定

第 9 章　はじめに

スポーツを強くするためには，得られたデータを客観的，かつ科学的に分析するアプローチが重要である．例えば，バレーボール女子日本代表では，体に心拍数を把握できるウェアラブルデバイスを装着し，その心拍数を監督が見ながら選手の調子を把握して選手交代を行っている．そこで本研究では，サッカー日本代表の戦術を一般的に用いられるベイジアンネットを用いて推定し，有効だった戦術を報告する．

第 10 章　提案手法

10.1　ベイジアンネット

問題領域の確率変数と変数間の関係をモデルで表したものである．事象の確率変数をノードとし，ノード間の相関や依存関係を有向辺で結んだ非循環有向グラフ DAG（Directed Acyclic Graph）で表す．リンク元のノードを親ノード、リンク先を子ノードと呼ぶ．結ばれたリンクの強さを各ノードの条件付き確率表 CPT（Conditional Probability Table）で表す．CPT の計算は DAG の構造を決定することで計算される．従って，ベイジアンネットを戦術推定に用いるためにはノードの決定，及び DAG の構築を行う必要がある．

10.2　ノードの決定，定義

表 14　モデルのノード名とその状態

変数	ノード名	状態	値域
X_1	エリア	ボールを奪取した位置	高，中，少
X_2	連携数	ラストプレーまでに経由した人数	少少，少，多，多多
X_3	進行度	エリアの横移動を 1，縦移動を攻撃方向によって±10 した値	大大，大，小，小小
X_4	ラストプレー	相手がボールに触れる直前のプレー	※後述
X_5	結果	シュートの是非	1,0

ノードはオフェンスの途中で変化する量である．エリア X_1はフィールドを縦に 5 分割，横に 3 分割したもので，攻撃の起点となる部分を示す，攻撃方向において右から 1~5，6~10，11~15 の値を割り当て，その値に応じて高，中，低の値を格納した．連携数 X_2はパスの回数を示す．進行度X_3はボールの移動を示しており，エリアによって移動の度合いを変化させた．ラストプレ

ーX_4は相手がボールを触る直前のプレーを示す．図 1 にエリア，表 2 にラストプレーの定義を示す．※フリーキックにはキックオフ，スローイン，コーナーキック，ゴールキックも含む

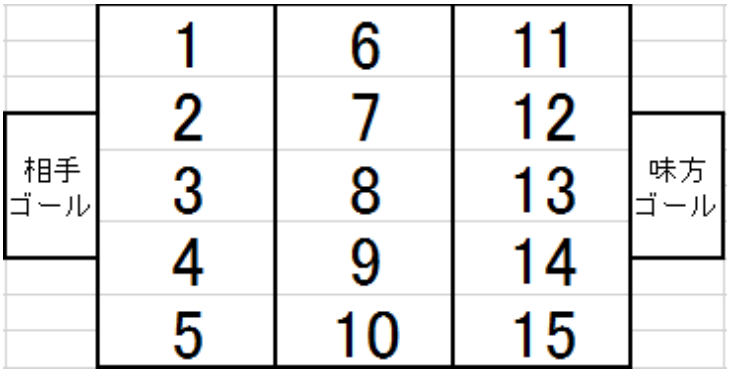

図 6　エリアの値の定義

表 15　ラストプレーの定義

X_4	プレー	定義
P_1	ショートパス	横に 2,縦に 1 エリア以内の範囲のパス
P_2	ドリブル	一人で 3 回以上触れるプレー
P_3	フリーキック	相手の反則から開始したプレー
P_4	ロングパス	ショートパス以上の距離のパス

変数 X_1，X_2，X_3は大きな値域の連続量となるので，それぞれのデータを非連続量の値に格納した．サッカーでは相手ゴールに近い場所を高い位置と表現するため，1~5 を高，6~10 を中，11~15 を低とした．同様に X_2，X_3はそれぞれ｛多多(10~)，多(6~9)，少(4~5)，少少(~3)｝，｛大大(23~)，大(15~22)，小(11~14)，小小(~10)｝として値域によって非連続の値を格納した．また，X_5はシュートの是非をブーリアンで示し，1 を true とした．

10.3　DAG の構築

DAG の構築には，R のパッケージの一つである deal を使用した．DAG の構造学習は，ネットワークスコア

S(G)=p(G,d)=p(d|G)p(G)

を利用して計算されるベイズファクター

S(G)/S(G*)=p(G|d)/p(G*|d)

によって評価される．ただし，DAG を構築するベイジアンネットワークモデルを G・G*，入力されたデータを d とする．このネットワークスコアの評価にはさまざまな方法が知られているが，deal ではよく知られている探索アルゴリズムの一つである欲ばり法が使われている[2]．

5 つの確率変数を考えたとき，得られるモデルは無数に存在するため，計算量を減らすために事前にあり得ない有向辺を削除する必要がある．時系列で考えると，ボールを奪取した際にエリアが決定し，相手選手がボールに触れた時点で連携数と進行度が決定する．同様にして確率変数間の時系列を考慮し，図 2 のような階層を定め，X_5からすべてのノード，X_4から X_1，X_2，X_3への有向辺などを削除した．

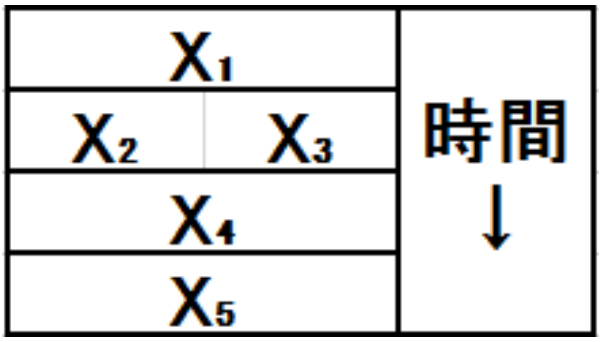

図 7　確率変数の階層構造

10.4 学習結果

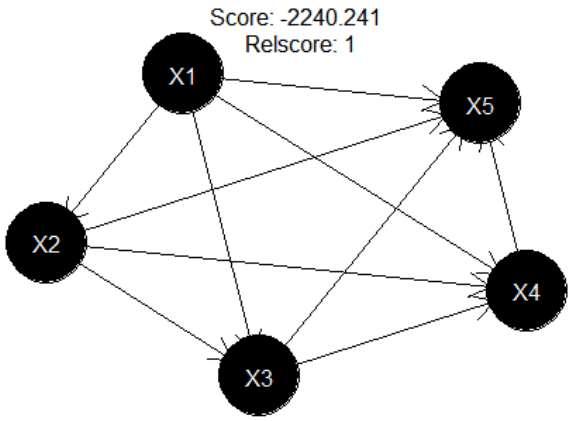

図 8　学習結果の DAG

図 3 に日本代表のデータについて求めたベイジアンモデルを示す．シュートできる確率を最大化する条件は P（X_1=低，X_2=多多，X_3=大，X_4=ショートパス|X_5=1）で 3.04%であった．次点以降，及びエリアごとの上位 10 確率は表 3，表 4 で示す．表 3 の総合的な順位で見てみると，ラストプレーの値はまちまちだが，エリアについては高い位置でボールを奪取したとき，ほとんどシュートできていないことが分かる．表 4 より最も確率が高い事象でも 0.33%と，非常に低い．

サッカー日本代表は細かくパスを回して相手をかく乱しているイメージが強かったが，2 位，4 位はラストプレーがドリブルであった．

エリアごとの特徴としては，エリア=高の時，3 事象すべて連携数が非常に少なく，進行度が小さい．すなわち，高い位置でボールを奪取したらなるべくボールを回さずに少ない人数で攻撃したほうがシュートできる確率が高い．この時取られている戦術はカウンターが考えられる．カウンターとは相手がディフェンスの陣形を組み立てる前に攻め崩してしまおうという戦術で，俗には速攻と呼ばれているものである．

エリア=中の時，上位 2 事象は連携数が多く，進行度が大きい．すなわち，中盤でボールを奪取したら，多めの人を経由し，シュートのために高い位置に移動するため，横移動は 5 回以上 12 回以内でラストプレーに繋げるとシュートできる確率が高い．この時取られている戦術はクロスが考えられる。クロスとは図 4 に示すようなプレーであり，図に分かる通り，クロスを上げてシュートするだけで横移動が 2 回ある．また，クロスをする前にはディフェンスのプレーヤーが横パスでどのサイドから攻めるか選択する段階がしばしば見られるため進行度が増えやすい戦術であるといえる．また，エリア=中のときのラストプレーの 1 位がロングパスであるのもクロスを上げている可能性が高いことを示唆する要因となっている．

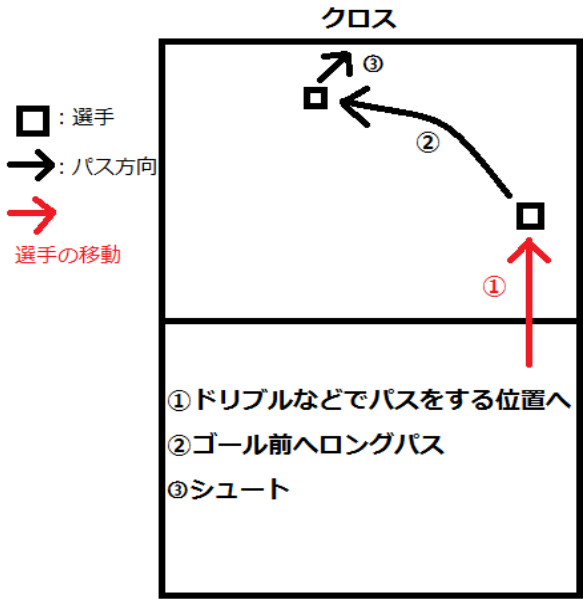

図 9 クロスの図解

エリア=低のとき，3 事象ともに連携数がとても多く，進行度も大きい，またはとても大きい．すなわち，低い位置でボールを奪取したらじっくりボールを回しながら横移動が 4 回以内だとシュートできる確率が高い．このとき取られている戦術はカウンターが考えられる．先述した通り，カウンターとは相手の守備の陣形が組み立てられる前に攻め崩そうとする戦術であるが，それゆえになるべく最短距離の縦移動で攻めることが多い．ラストプレーがショートパス，ドリブルであるのもカウンターで攻めている可能性が高いことを示唆している．

表 16　シュートできる確率が高い 10 事象

順位	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
確率	0.0304	0.0284	0.0282	0.0269	0.0256	0.0255	0.0221	0.0212	0.0206	0.0179
エリア	低	低	中	中	中	中	低	低	低	低
連携数	多多	多多	多	多	多	多	多多	多多	多多	多多
進行度	大	大	大	大	大大	大	大大	大大	大大	大
ラストプレー	S パス	ドリブル	L パス	ドリブル	S パス	S パス	L パス	ドリブル	フリーキック	フリーキック

表 17　エリアごとの P(結果=YES)を大きくする要因上位 3 事象

順位	確率	連携数	進行度	ラストプレー
高 1 位	0.0033	少少	小	フリーキック
高 2 位	0.0031	少少	小小	ショートパス
高 3 位	0.0023	少少	小	ドリブル
中 1 位	0.0282	多	大	ロングパス
中 2 位	0.0269	多	大	ドリブル
中 3 位	0.0256	多	大大	ショートパス
低 1 位	0.0304	多多	大	ショートパス
低 2 位	0.0284	多多	大	ドリブル
低 3 位	0.0221	多多	大大	ロングパス
高平均	0.0011	—	—	—
中平均	0.0078	—	—	—
低平均	0.0068	—	—	—

第 11 章　おわりに

本結果では，ボールを奪取するエリアが低いほどシュートできる確率が高く，その中でも多くの人数を経由することでシュートをする確率が高いことを明らかにした．しかし，本研究はラストプレーにのみ着目したので，その間の決定的な攻撃の構造の発見には至らなかった．今後は，その対策として，(1) 1 プレーごとにショートパス，ロングパス、ドリブル，フリーキックの回数を記録する，(2) ボールの動きに着目した確率変数を設定する，(3) 日本代表だけではデータ数が少ないので J リーグなどリーグ単位でデータを入力する，(4) 1 プレーを評価できる値を設定する，といったことを検討する．

また，データ入力にあたり，TV 中継の映像では視点が移動するためエリアやラストプレーの判断が困難であった場面がしばしばあったため，視点を固定させた映像データも求められる．

参考文献

[1] 上原司，荒井秀一，MLB 詳細スコアデータから学習した試合構成群間における確率的因果構造に基づく野球選手の投球戦術推定，電子情報通信学会 IEICE，pp. 1-5，2012 年

[2] 豊田秀樹，データマイニング入門，東京図書，pp. 209-240，2014 年