顧客のセグメンテーションと商品のスコアリングによる購買予測

伊藤　孝太朗，澤邊　剛，保坂　桂佑，松下　亮祐，雪島　正敏

1. はじめに

1.1 本論文の構成

経営科学系研究部会連合協議会主催の平成 25 年度データ解析コンペティション課題部門には，EC サイトの閲覧，購買履歴データが提供された．課題は，顧客別に予測期間の最初の購買商品を予測することであり，予測精度に基づく得点により順位付けが行われる．我々は「remember3 丁目」というチーム名で本コンペティションに参加し，優秀賞（順位は 2 位）を受賞した．本論文ではまず我々がどのように分析を進めていったか紹介したのち，我々が実際に用いた予測手法を大きく 3 つの手順に分けて説明する．最後に，分析の取り組みの反省点を述べる．

1.2 コンペティションについて

データの提供元となった EC サイトは主に衣服類を扱っている．本コンペティションで提供されたデータには，顧客の性別や商品のブランドなどの属性情報のほか，2011 年 9 月から 2013 年 4 月までの購買履歴，閲覧履歴が含まれている．これらのデータをもとに翌月 2013 年 5 月の各顧客の最初の購買商品を予測せよというのが課題の内容であり，コンペ参加者は各顧客に対し，予測商品を 6 つ挙げる．各商品には商品を最も細かい単位で一意に区別する商品詳細 ID と，サイズ，色の違いを無視した商品 ID とが付けられている．商品 ID と商品詳細 ID，それぞれの予測的中数によって予測手法の優劣が決まる．

2. 分析の手順

本コンペティションは，精度の高い予測モデルを作成することが目的であるが，我々は予測モデル作成の前段階として，データの概略を把握するためにかなりの時間を事前分析に費やした．この節では事前分析，予測モデル作成のそれぞれのフェーズについて，どのように分析を進めていったか紹介する．

事前分析では，各量の頻度集計といった基本的な集計のほかに，予測モデル作成の際，どのような変数が有用な説明変数になりうるかということを中心に調べた．例えば，商品の閲覧回数など商品の購買と関係がある予想される変数に対して，具体的に変数が変化したとき，商品の購買確率や購買時期などがどの程度変化するかを調べた．このフェーズでは，隔週でミーティングを設け，各メンバーの集計結果を共有するとともに次回までの集計項目の分担を決定した．コンペティション開始から中間スコア提出までの期間をこのフェーズに充てた．

中間スコア提出時点から，本格的な予測モデル検討のフェーズに入った．後述するように，事前分析で顧客を 2 つのグループに分けて予測することが有用であると示唆されたため，メンバーの分担を決め，それぞれ担当の顧客グループに対する予測モデル構築に集中した．このように分担を決めたことにより，同じ対象に対して分析が発散，重複せずに，スムーズに分析を進めることができたように思う．

3. 予測手法の大枠

本節では，我々が実際に用いた予測手法の大枠を述べる．購買履歴には商品 ID，商品詳細 ID の情報が両方存在するが，閲覧履歴には，商品 ID の情報のみ存在する．また，商品 ID の粒度を細かくしたものが商品詳細 ID なので，商品 ID の予測が的中しない限り商品詳細 ID の予測は的中しない．そこで，本手法では，商品の予測を商品 ID の予測と商品詳細 ID の予測との二段階に分けて行った．まず第一段階として各顧客に対し商品 ID を 6 つ予測し，次に第二段階として各顧客に対して予測された商品 ID の各々に対し，予測商品詳細 ID を 1 つ決定した（図 1）．

第一段階においては，予測期間の直近 1 カ月に閲覧

いとう　こうたろう，さわべ　つよし，ほさか　けいすけ，まつした　りょうすけ，ゆきしま　まさとし
（株）NTT データ数理システム
〒 160–0016 東京都新宿区信濃町 35 信濃町煉瓦館 1 階

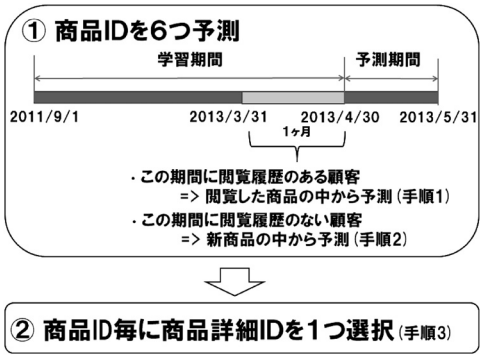

図 1　予測手法の大枠

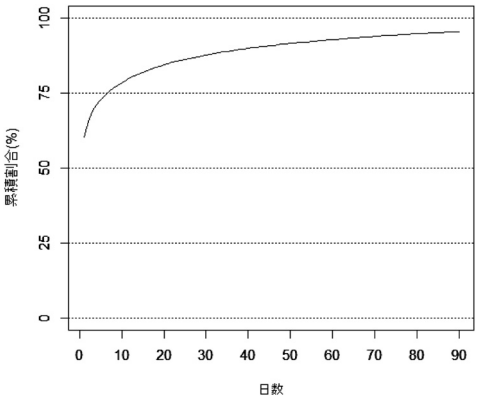

図 2　顧客が商品を購買した時点における，その顧客がその商品を初めて閲覧してからの経過日数の累積度数分布．縦軸は累積度数を全購買数で割った値．

履歴があるかないかで顧客を 2 つの群に分け（図 1），それぞれの群の顧客に対して異なる手法を用いて予測を行った．具体的には，直近 1 カ月に閲覧履歴がある顧客には主に閲覧履歴にある商品の中から予測商品を選び，そうでない顧客には「新商品」の中から予測商品を選んで予測を行った．閲覧商品の中からの予測には目的変数の偏りを考慮したランダムフォレスト [1] を，「新商品」からの予測には商品属性に基づく商品のスコアリングを用いた．

なお，提供されたデータからは発売は既にされているが購買はされていない商品と新しく発売された商品とを明確に区別することはできない．そのため，我々は発売は既にされているが購買されなかった商品も「新商品」として扱った．つまり，「ある月の新商品=その月より前の購買履歴に存在しない全商品」とした．本論文でも「新商品」で既に発売されているが購買はされていない商品と新しく発売された商品両方を表すとする．

また，第一段階において前述のように顧客を分類をした理由は，顧客が商品を購買した時点における，その顧客が対象商品を初めて閲覧してからの経過日数が短いことが多いためである．

図 2 が示すように，顧客が将来購買することになる商品を初めて閲覧してから購買するまでの日数は，30 日以内のものが全体の購買の 90%近くを占めている．この事実は，顧客が予測期間の直前に閲覧した商品があればその中から予測商品を選ぶという手法の妥当性を示唆している一方，閲覧履歴が予測期間の何カ月も前のものしかない顧客に対しては，閲覧商品の中から商品を選んで予測するのは有効ではないことも示唆しており，前述のような顧客の分類を行うことへの動機づけとなっている．

また，前述のとおり，閲覧履歴には顧客が閲覧した商品 ID の情報があるが，商品詳細 ID の情報はない．そのため，第二段階の商品詳細 ID の予測においては，両方の群の顧客に対して購買履歴をもとにした同一の手法を用いた．ここでは二項ソフトクラスタリングという手法を用いた [2]．

本論文では，直近の閲覧履歴のある顧客に対する商品 ID の予測を手順 1，直近の閲覧履歴のない顧客に対する商品 ID の予測を手順 2，商品詳細 ID の予測を手順 3，と呼ぶ．以下，具体的な予測手法について，手順 1，手順 2，手順 3 の順に説明する．

4. 予測手法の詳細

4.1 手順 1 ：直近の閲覧履歴のある顧客に対する予測

直近 1 カ月に閲覧履歴のある顧客に対して閲覧商品から予測する商品を決定する手法として，ランダムフォレストを用いた．購買商品を直接予測するのではなく，閲覧商品の購買有無を予測するモデルを作成した．また，モデル作成時には目的変数の偏りを考慮して学習用データのサンプリングを行った．本節ではこの手法について説明する．

閲覧のあった商品に限定しても，商品 ID は 7,000 個以上存在するため，直接目的変数にするには適さない．そこで，前述のように，ある月に閲覧のあった顧客と商品の組に対して，商品が顧客の翌月の最初の購買商品となるか否かを目的変数としてランダムフォレストの学習を行った．なお，ランダムフォレストのアルゴリズム内で生成する決定木の数は 500 とした．そして，顧客ごとにランダムフォレストにより求まる購買確率の高い順に商品 ID を予測した．

表 1　ランダムフォレストの重要度の高い説明変数の抜粋と重要度

説明変数	重要度
対象の商品は，顧客の対象の月における最後の閲覧商品か否か	3680.2
顧客の，対象の商品の閲覧回数	3206.7
対象の月に顧客が閲覧した商品の種類数	2922.2
対象の月の顧客の閲覧回数の合計	2668.0
対象の商品が顧客のお気に入りショップのものか	1570.0
顧客の購買履歴のうち対象の商品と同じブランドを購買した割合	1485.5
顧客の対象の月の最後の閲覧から翌月までの時間	1396.7
対象の商品の年齢と，顧客が購買した商品の平均年齢との差	1282.4
顧客が購買した商品の年齢の中央値	1197.6
対象の商品の対象の月における女性による閲覧数	1186.1

説明変数は，計 21 個を作成した．ランダムフォレストは大量の決定木を学習させるが，各説明変数の分割による Gini 係数の減少度をすべての木で平均して，説明変数の重要度を算出することができる．その重要度が高い説明変数上位 10 個が表 1 である．ここで表 1 中の「商品の年齢」とは，初めて商品が購買された日から対象の顧客に閲覧された日までの日数である．

次に，目的変数の偏りを考慮したデータのサンプリング手法について説明する．上記のようなデータの特徴として，目的変数の偏りが大きい，つまり顧客が閲覧したうえ購買した商品（正例）よりも購買しなかった商品（負例）のほうが圧倒的に数が多いということが挙げられる．実際のデータでは正例と負例の比が 1:153 ほどであった．このようなデータのクラス分類をする場合の有用な手法として，サンプリングにより正例と負例のバランスをとることが知られている [3, 4]．本コンペティションでもサンプリングによる手法を扱った．

さまざまな正例と負例のバランスを試した結果，コンペティションの課題と同様に商品 ID を 6 個予測した場合の的中数が最も多くなったのは，学習用データの正例を 1.5 倍に over sampling し，全体として正例と負例の比が 1:10 になるように負例を under sampling した場合であった．この手法によりサンプリングした場合，およびサンプリングを行わずに正例と負例の比が 1:153 のままでランダムフォレストを学習した場合の商品 ID の的中数を表にしたものが表 2 である．表 2 では，学習用データとして，分析用マシンのメモリの都合上 2012 年 10 月から 2013 年 1 月までの 4 カ月間のデータをレコード数が 6 分の 1 になるようにランダムサンプリングしたものを用い，2013 年 2 月から 2013 年 3 月のデータに対して，商品 ID の的中数を算出した．上記のように正例と負例の比を調整することで，6 個商品を予測した場合で商品 ID の的中数は 4%ほど向上した．

表 2　商品 ID の的中数

予測する個数	ランダムフォレストによる購買確率の高い順に予測した場合の的中数		顧客ごとの閲覧回数の多い順に予測した場合の的中数
	正例：負例＝1:10	正例：負例＝1:153	
1	352	361	317
2	531	528	481
3	651	630	584
4	750	715	687
5	824	787	772
6	889	855	842
7	953	904	908
8	990	951	948
9	1029	994	995
10	1065	1031	1035

4.2　手順 2：直近の閲覧履歴がない顧客への予測

第 3 節で述べたように，予測期間の直近 1 カ月に閲覧履歴のない顧客に対しては新商品の中から予測商品を選んで，商品 ID の予測を行った．予測においては，各顧客に対し，新商品ごとのスコアを商品属性をもとに計算し，スコアの大きいものを予測商品とした．顧客 X に対して商品 ID の予測を行うときの，スコア算出式を含めた具体的な予測手順は以下のとおりである．

1. すべての新商品 A に対し，新商品 A のスコアを以下の式で計算する．

（顧客 X に対する商品 A のスコア）
＝（顧客 X が過去に商品 A と同じ小分類の商品を購買した回数）
＋（顧客 X が過去に商品 A と同じショップの商品を購買した回数）
＋（顧客 X が過去に商品 A と同じブランドの商品を購買した回数）
$-\alpha \times$（商品 A と三属性すべてが同じ新商品の数）．　(1)

ここで，三属性とは小分類，ショップ，ブランドの 3 つの商品属性を指す．また，α は後述の方法によりチューニングされるパラメータである．

2. 1. で算出したスコアの大きい上位 6 つの商品 ID を予測商品 ID とする．ただし，同スコアの新商品が複数ある場合には，商品 ID が大きいものを優先して予測商品 ID とする．

以下で，この手法の意図するところを述べる．

予測商品候補を新商品に絞ったのは次のような理由からである．本コンペティションで提供されたデータの特徴として，商品寿命が概して短いことが挙げられる．したがって，過去に購買された商品は予測期間には購買されない可能性が高く，過去に購買された商品よりも新商品のほうが予測期間に購買される可能性が高いことが期待できる．実際，新商品は提供データ中の全商品数の 5%程度しかない一方，1 カ月の全購買数の約 20%を占めており，ある商品が新商品であるということが，その商品が予測期間に購買される確率に正の影響を与えている．

新商品の中から選んで予測する場合には，商品の属性をもとにした予測，いわゆる，コンテンツベースのフィルタリングを行うことが必要となる．本手法では，商品属性の中から，色・サイズのような商品詳細 ID に紐づくもの，商品大分類のように別の属性を集約することによって作られるものを除いた，商品小分類，ショップ，ブランドの 3 つの属性を用いた．

スコアの算出式 (1) は，商品 ID が当たる確率を向上させるという目的のもと設計されている．商品 ID が当たる確率は商品の三属性が当たる確率と商品の三属性が当たったときに商品 ID が当たる条件付き確率との積によって決まる．したがって，この積に現れる 2 つの因子の両方の値をバランス良く向上させることが，商品 ID が当たる確率の向上には重要である．これら 2 つの因子のうち前者，商品の三属性が当たる確率を向上させるために導入したのが，(1) の右辺の始めの 3 つの項である．これらの項は，顧客は過去に購買した商品と同じ小分類，ショップ，ブランドの商品を購買しやすいという仮定のもと導入した．式 (1) において $\alpha = 0$ としたときには，これらの項のみでスコアが計算されるが，そのときの商品 ID の的中率はランダムに予測した場合の的中率よりも高く，この仮定の妥当性が示唆される．

一方，(1) の右辺の最後の項は，後者の因子，すなわち商品の三属性が当たったときに商品 ID が当たる条件付き確率を向上させるために導入した罰則項である．新商品によって，自身と三属性がすべて同じである新商品の個数にはばらつきがあり，この個数が大きい新商品は三属性が当たったとしても商品 ID は当たりにくいため，そのような新商品に大きな罰則を与えるようにしてある．この罰則項の中に現れる係数 α は，商品の三属性が当たる確率と商品の三属性が当たったときに商品 ID が当たる条件付き確率とのバランスを調節するためのものである．係数 α の値は，男女ごとにいくつかの値を試し，最も予測精度が良い値を採用した．

また，同スコアの新商品が複数ある場合に商品 ID が大きいものを優先して予測商品 ID とするという手法も，商品の三属性が当たったときに商品 ID が当たる条件付き確率を大きくするためのものである．商品 ID は商品の購買時期と相関があり，商品 ID の大きいもののほうがより新しいと考えられる．本手法では提供データのうち購買履歴のない商品をすべて新商品として扱ったので，予測の際，既に発売されているが一度も購買されていない商品を選んでしまうリスクがある．そこで，商品 ID の大きいものを優先すればこのリスクが軽減することが予想できる．実際，同スコアのものが複数ある場合に，ランダムに選ぶ，商品 ID の大きい順に選ぶ，商品 ID の小さい順に選ぶ，という 3 つの手法を試した結果，商品 ID の大きい順に選んだ場合が最も予測精度が高かった．

4.3 手順 3：商品詳細 ID の予測

この節では，手順 1，2 により商品 ID を決定した後，商品 ID ごとに商品詳細 ID を予測する手法について説明する．

商品詳細 ID は商品 ID と商品のサイズ，色，により決定される．本手法ではまず，商品詳細 ID に対して顧客のサイズと色の「好み」を算出する．そして，これらを一定の割合で足し合わせたものをスコアとし，商品 ID ごとにスコアの最も高い商品詳細 ID を予測とした．サイズや色には表記ゆれがあるが，手動で表記ゆれを補正する代わりに，表記ゆれを補正と「好み」の算出を同時に行う手法として二項ソフトクラスタリングを用いた．以下，表記ゆれによって生じる問題と，二項ソフトクラスタリングについて説明する．

データ提供元の EC サイトでは主に衣服などを取り扱っているので，一度「大きい」サイズを購買した顧客は，別の商品を購買する際も，「大きい」サイズを好むと考えられる．しかし，ここで問題になるのは，商品によってサイズや色のフォーマットが異なることである．サイズのフォーマットが「S，M，L」という商品もあれば「SIZE: S，SIZE: M，SIZE: L」と表される商品もある．例えば，顧客のサイズの「好み」を，同じサイズの商品の購買履歴の有無で測る場合，次のような問題が生じる．つまり，サイズが「S，M，L」で表される商品の購買履歴がない場合，過去に「SIZE: L」サイズの商品を購買したことのある顧客は「S」サイズの商品よりも「L」サイズの商品を好む，ということが表現できない．さらに，サイズの種類は 1,000 以上

に上り，サイズのフォーマットの名寄せを手動で行い「大きい」「小さい」といった分類をするのは困難である．そこで我々は，これらの困難を回避して顧客のサイズ（色）の「好み」を算出する手法として，二項ソフトクラスタリングを用いた．二項ソフトクラスタリングは，株式会社 NTT データ数理システムの Visual Mining Studio の機能である．

二項ソフトクラスタリングでは，2 つの属性 X_k, Y_l の出現頻度をもとに，共起確率（同時に出現する確率）$P(X_k, Y_l)$ を算出する．さらに，X_k, Y_l が属するクラスタを表す潜在変数 Z_m を導入する．各属性がクラスタ Z_m に属すると仮定したとき，X_k が起こる確率を $P(X_k|Z_m)$，Y_l が起こる確率を $P(Y_l|Z_m)$ とする．これらとクラスタが出現する確率 $P(Z_m)$ により，$P(X_k, Y_l)$ を

$$P(X_k, Y_l) = \sum_m P(X_k|Z_m)P(Y_l|Z_m)P(Z_m) \quad (2)$$

と表現する手法が二項ソフトクラスタリングである．

今回は X_k を顧客，Y_l をサイズ（色）として，顧客ごとの対象のサイズ（色）の商品の購買回数をもとに，$P(X_k, Y_l)$ を算出した．クラスタの数は，サイズの場合は 4，色の場合は 10 とした．そして，顧客を固定したときサイズ（色）が出現する確率 $P(Y_l|X_k)$ を顧客のサイズや色の「好み」と解釈した．(2) より $P(Y_l|X_k)$ は次のように書ける．

$$P(Y_l|X_k) = \sum_m P(Y_l|Z_m)P(Z_m|X_k) \quad (3)$$

$P(Z_m|X_k)$ は顧客がクラスタに属する確率，$P(Y_l|Z_m)$ はクラスタから各サイズ（色）が出現する確率である．

二項ソフトクラスタリングでは，サイズや色に表記ゆれがあっても，購買傾向が似ている顧客に購買された商品のサイズや色は，同じクラスタから出現する確率が高くなり，前述のような問題は起こりにくい．実際，クラスタごとに $P(Y_l|Z_m)$ の高い順にサイズ（色）を並べたものは，名寄せを行っているように見えることも観察でき，クラスタリングにより表記ゆれが補正されていると考えられる．クラスタごとに $P(Y_l|Z_m)$ の高いサイズを並べたものの抜粋が表 3 になる．各クラスタに，同じような大きさを表すと思われるサイズが並んでいることが観察できる．

本手法の有用性の検討は，2012 年 10 月 1 日から 2013 年 1 月 31 日までの 4 カ月間のデータを学習用データ，2013 年 2 月から 2013 年 3 月のデータを検証用データとして行った．商品 ID が的中したとき商品詳細 ID が的中する確率は，ランダムに商品詳細 ID を予測した場合は 22%程度だが，本手法を用いると 62%ほどに向上した．

表 3　各クラスタが生成するサイズ

クラスタ 1	クラスタ 2	クラスタ 3	・・・
SMALL	LARGE	ONE SIZE	・・・
1	X-LARGE	フリー	・・・
X-SMALL	3	フリー	・・・
0	43	40	・・・
44	4	F	・・・
・・・	・・・	・・・	・・・

5. おわりに

本コンペティションにおける我々の最終的な成績は 2 位であった．最後に，手順 1，手順 2，手順 3 の順に分析における反省点を考察する．

手順 1 では，直近に閲覧履歴のある顧客に対してランダムフォレストを用いたモデル構築を行った．実際の分析での反省点としては，顧客の予測期間の直近 1 カ月の閲覧商品が 6 個に満たない場合の予測が上げられる．今回我々は，閲覧商品が 6 個に満たない顧客に対しては，予測商品が 6 個になるまで，男女別の人気商品を予測商品としたが，この手法の十分な検討を行えなかった．閲覧数が 6 個に満たなかった顧客は，直近に閲覧履歴のある顧客の 33%にのぼるため，この手法の検討を行うことにより，より精度が向上すると考えられる．直近の閲覧履歴のない顧客に対しては，人気商品を予測するよりも，新商品を予測する方が精度が良かったため，改善案として，人気商品の代わりに新商品を予測する方法が考えられる．さらに，商品詳細 ID の的中率を向上させるために，同じ商品 ID で違う商品詳細 ID の商品を複数予測する方法なども考えられる．また，ランダムフォレスト以外の手法を十分に検討できなかったのも反省点の一つである．

手順 2 では，新商品をスコアリングし，直近に閲覧履歴のない顧客に対して予測を行った．購買履歴がない商品，特に新商品を予測することができる点が，本手法の優れた点である．反省点としては，予測手法の頑健性を確保することに力をより注ぐべきであったという点が挙げられる．例えば，式 (1) における罰則項の係数を決定する際，2013 年 2 月から 2013 年 3 月の購買の予測精度が最大となるように決定した．したがってモデルが過学習していると考えられ，今回決定した係数が予測データに対して必ずしも最適でないことが想像される．異なる複数の期間のデータを用いて検証を行うことにより，頑健性が向上すると考えられる．

手順 3 では，二項ソフトクラスタリングを用いて商

品詳細 ID を予測した．購買履歴さえあれば，各商品属性に対する顧客の「好み」を算出でき，副産物として，似かよった属性をまとめた表が得られる点が本手法の優れた点である．実際の分析での反省点としては，予測手法を検証する仕組みを十分に整えられなかった点が挙げられる．二項ソフトクラスタリングにおける潜在変数の数，スコア算出の際のサイズと色の「好み」を足し合わせる割合が，本手法のパラメータであるが，実際の分析ではいくつかのパラメータを散発的に調べたのみである．そこで，サイズと色の「好み」を説明変数として，商品詳細 ID の購買有無を予測するモデルを新たに構築するなどの改善手法が考えられる．

参考文献

[1] L. Breiman, “Random forests,” *Machine Learning*, **45**, 5–32, 2001.

[2] T. Hoffman, “Probabilistic latent semantic analysis,” *Proceedings of the Fifteenth Conference Annual Conference on Uncertainty in Artificial Intelligence*, 289–296, 1999.

[3] C. Ling and C. Li, “Data mining for direct marketing problems and solutions,” *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)*, 73–79, 1998.

[4] M. Kubat and S. Matwin, “Addressing the curse of imbalanced training sets: OneSided selection,” *Proceedings of the Fourteenth International Conference on Machine Learning*, 179–186, 1997.

