O-008

大規模食事記録データの栄養価クラスタリングに基づく食習慣ベクトル Bag-of-Foods とその検証

Bag-of-Foods: Vectorizing Diet Preference based on Nutrition Clustering of Large Foodlogging Data

合田 悠治† 天野 宗佑† 山肩 洋子† 相澤 清晴†
Yuji Goda Sosuke Amano Yoko Yamakata Kiyoharu Aizawa

1. まえがき

食事と健康状態には密接な関わりがあり，食習慣の把握が健康診断の上で重要であるのは周知のとおりである．一方，食事は嗜好品の側面も持っており，健康管理の観点では不要な食事をとることも少なくない．このように食習慣は個人の健康状態・性格といった要素が複雑に絡んで形成されるものであり，その分析は健康診断に限らず世間的な消費傾向の調査などでも重要である．近年，携帯端末が普及し食事ログの支援ツール [1] も現れたことで，従来よりも大規模な食事記録データが得られるようになった．以上の背景を踏まえ，本稿では大規模食事記録データを用いた食習慣の分析及びベクトル化について検討する．まず，栄養価に基づいた食習慣の分類が直観に沿った結果を示すことを確認した．これを受け，食習慣のベクトル化手法として Bag-of-Foods を提案し，これが食習慣の表現として妥当であるか検証した．

2. 関連研究

本研究では食習慣に注目しているが，ヘルスケアの文脈では食事記録に運動量，喫煙量といったデータを加えて，より広いスコープのライフスタイル分析を行っている研究が一般的である．もっとも，肥満や特定の病気などを対象にその要因を生活習慣に見出そうとする研究は古くから存在するが，データドリブンに生活習慣の分析を行う研究は比較的少ない．

ライフスタイル分析を統計的に行う研究として，Kesse-Guyot らによる研究 [2] や，Patino-Alonso らによる研究 [3] などが挙げられる．これらは食材ごとの摂取頻度やアルコール摂取量，運動量，喫煙量，血圧などをライフスタイルデータとして分析したもので，不健康要因となるライフスタイルを統計的に発見することを目的としたものが多い．特に，Patino-Alonso らの研究 [3] ではライフスタイルのクラスタリングが行われているが，野菜の摂取量と果物の摂取量に統計上大きな意味的差異がない一方で，アルコールの摂取量は野菜摂取量とほとんど逆の意味を持つことが結果として示され，各ライフスタイルは一般に健康的とされる要因と不健康とされる要因とのバランスによって健康クラスタ・不健康クラスタに分かれるという直感的に理解しやすい結論が得られている．

このようにライフスタイル分析は既存の研究でもある程度明瞭な結果が出されているが，食習慣の分析としては不足の部分もある．例えば上に挙げた研究では，食事記録に Food Frequency Questionnaire (FFQ) [4] を用いている．FFQ は決められた食材ごとに何をどの頻度で食べるか記録したもので，健康状態に影響のありそうな食材の摂取量のみ抽出する目的では便利だが，これはどの食材が健康上重要かという事前知識を前提としており，また時間的な情報も持つ食事ログと比べるとデータとしては失われている情報が多いと思われる．

大規模食事記録を扱った研究としては，椿田らの研究 [5] が挙げられる．これは協調フィルタリングによって疎な食事記録から密な食事記録を予測するという研究で，協調フィルタリング手法には Matrix Factorization (MF) を使っている．MF は次元削減を行い暗黙的にユーザ特徴を作る手法で，その点では食習慣の特徴量生成を目的にする本研究とも通ずる点がある．ただし，こちらは食事記録そのものを復元することが目的であるので栄養価には触れられていない．

3. 提案手法

3.1. データセット

本研究では，2013 年 7 月から 2017 年 3 月までの期間に食事ログツールである FoodLog[1] へ登録されたデータをもとに実験を行う．FoodLog データの一件一件は，ユーザ ID・登録時刻・料理名・量…のように食品ごとに登録されており，デフォルトで用意された 1,870 種類の料理名については一人前当たりの標準栄養価が紐づいている．ユーザが新たに料理名を追加する機能もあるが，本研究は摂取栄養価に基づいた分析が目的なので，ユーザ追加料理名の食品はできるだけ含まれないことが望ましい．

また，摂取栄養価を計るならば一定期間中に食べた食品をまとめて考慮に入れる必要がある．食事における時間的な単位としては一食ごと，一日ごと，一週間ごとなどが考えられるが，FoodLog にユーザの全食事の記録が登録されている保証がなく，ユーザから登録のない期間にユーザが何も食べていないのか食べたが登録をしていないのかの判断がつかないため，一食を単位にするのが適切である．一食の定義については，FoodLog が一食を一枚の写真に収めて一度に登録することをユーザに推奨しており，その場合の食事記録はすべて同時刻のレコードを持つことから，同時刻に記録された食品群を一食と扱うこととした．

以上のことを踏まえ本研究での実験は，FoodLog に登録されているデータの 9 割以上が標準メニューのもので，かつその件数が 1,000 件以上である 157 名の記録 (FLD541K) を対象に行った．登録されているレコードは 541,712 件で，187,609 食分のデータとなる．

†東京大学 情報理工学系研究科 電子情報学専攻

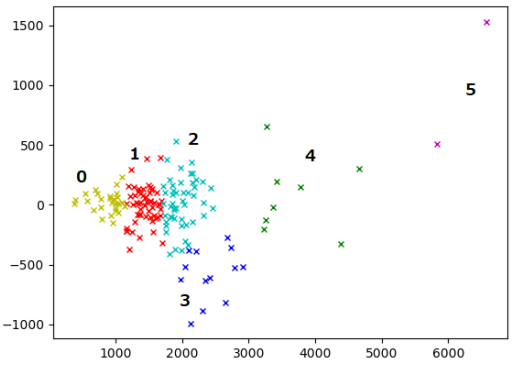

図 1: 一食当たりの平均摂取栄養価による K-means クラスタリングの結果

表 1: 図 1 における各クラスタでの一食の献立の例

クラスタ	献立
0	磯辺もち
	焼きとり, 焼酎
1	おでん, ゆで卵, ご飯
	ほるさめのスープ, ご飯, 納豆
2	うなぎ丼, アイスクリーム
	ヨーグルト, ブラックコーヒー, ハムサンド, 野菜ジュース
3	バナナ, ブラックコーヒー, フルーツ盛り合わせ, ヨーグルト, グリーンサラダ, ジャムトースト
	生野菜サラダ, カレーライス, ほうれん草の煮びたし
4	野菜炒め, ご飯, 焼き魚, 味噌汁
	ご飯, ミートローフ, 中華風スープ, おでん, グリーンサラダ
5	ご飯, きんぴら, 味噌汁, 鶏のから揚げ, 昆布としいたけの佃煮, 焼き魚, キャベツの千切り
	串揚げ, キャベツの千切り, 煮卵, ご飯, あんドーナツ, 焼きビーフン, きんぴら, 味噌汁, ほうれん草のごま和え, えびのチリソース煮

3.2. 予備実験：ユーザ分類

まず，食習慣の分析において栄養価ベースのアプローチが妥当であることを確認するために一食当たりの平均摂取栄養価によるユーザ分類を行った．

はじめに，ユーザごとに一食当たり摂取している栄養価の平均を算出した．個々の食事に対する栄養価は [6] に挙げられている 31 項目‡で表す．各項目の数値は [6] で挙げられる一日の推奨摂取量に対する摂取量の割合で示され単位はない．予備実験では，ユーザごとの平均摂取栄養価をそのままクラスタリングにかけた．

対象ユーザ 157 名に対し K-means によってクラスタリングを行うと図 1 の結果が得られた．なお，ここでは図示のために次元を落としてあるが，クラスタリングの段階では次元削減は行っていない．

図 1 でそれぞれのクラスタに属するユーザについて典型的と思われる一食の献立の例を表 1 に示す．

図 1 と表 1 を比較すると，図 1 において横軸の負の側に集まるユーザは一食あたりに食べている量が少なく，正の側に寄るほど食べている量が増えることが分かる．また，縦軸の負側に外れていて目立つクラスタ 3 の一食の献立を見ると，比較的一食に占める野菜の量が多いことが分かる．

以上より，摂取栄養価をそのままクラスタリングにかけるだけでも，小食の群，過食の群，野菜偏重群など，ある程度直感的に正しいユーザ分類ができることが分かった．この結果を踏まえ，食習慣のベクトル化手法として一食ごとの栄養価に基づく Bag-of-Foods を提案する．

3.3.Bag-of-Foods

Bag-of-Words では文章中にどの単語がどの程度含まれるかということを見るが，本手法では Word に相当するものとして食事の属性を使う．食事の属性を得るためには事前に食事の分類を行う必要があるが，3.2 章で直感に沿う分類が行えたことから 3.2 章と同様の方法で一食ごとの栄養価をクラスタリングにかけることとし，どの属性の食事をどの割合で食べているかという情報を個人の食習慣ベクトルとして扱う手法を提案する．

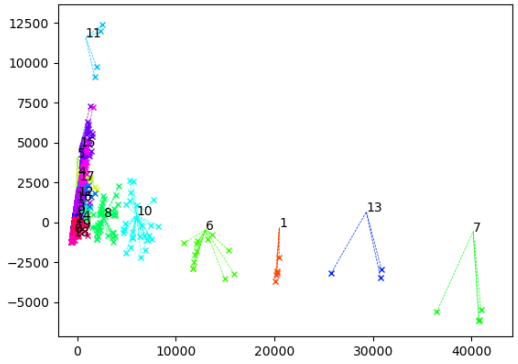

図 2: 栄養価に基づいた食事クラスタリング (K-means)

食事属性を得るためのクラスタリングでは，一食当たりの栄養価をクラスタリングにかけた．適切なクラスタ数の議論は 4.3 節で行うが，ここでは 20 に設定した．

クラスタリング結果を図 2 に示す．図の左側に密集して直線状に並ぶ群と，そこから右側へずれる群が存在するが，左に密集している群は図の上側へ行くほど全般に食べる量が増していることが分かった．一方，横に伸びる群は一食当たりの量は特徴的でないが，野菜偏重・塩分過多など栄養バランスに偏りのある食事が分類されており食事の属性付けとして十分な結果が得られた．

こうして得た一食ごとの属性をもとに Bag-of-Words の手法でユーザ食習慣のベクトル化を行う．すなわち，

‡エネルギー, タンパク質, 脂質, 炭水化物, 塩分相当量, カリウム, カルシウム, マグネシウム, リン, 鉄, 亜鉛, 銅, マンガン, ヨウ素, セレン, クロム, モリブデン, ビタミン A, ビタミンD, ビタミン E, ビタミンK, ビタミンＢ１, ビタミンＢ２, ビタミンＢ６, ビタミンＢ１２, 葉酸, パントテン酸, ビタミンC, ビオチン, ナイアシン, 食物繊維

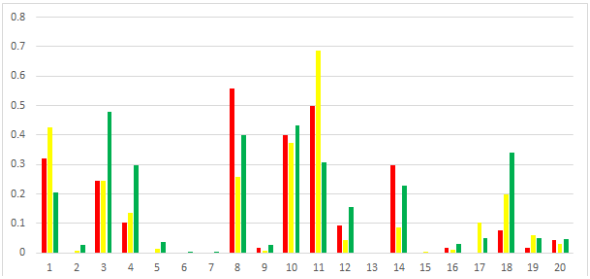

図 3: Bag-of-Foods 特徴量の例 (20 次元)

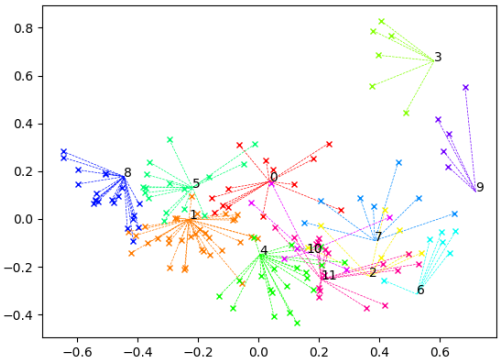

図 4: Bag-of-Foods に基づくユーザのクラスタリング (K-means)

ユーザごとにどの属性の食事を何回食べたかというベクトルを作り，それを正規化して食習慣ベクトル Bag-of-Foods とする．

図 3 に Bag-of-Foods 特徴量の例を示す．次元数は食事分類でのクラスタ数と同じ 20 となる．

4. 評価

食習慣は同一人物のものでも時間とともに移り変わりがあり，また食習慣同士の近さを絶対的に定義することはできないため，現状食習慣ベクトルの定量的評価を行うことが困難である．本稿では，まず提案手法を定性的に評価する．

4.1. 平均摂取栄養価クラスタリングとの比較

3.2 章ではユーザの平均的な食事傾向を分類できた．提案手法がまず平均的な食事傾向をどの程度表現できているかを確認するため，提案手法でユーザ 157 人をベクトル化しクラスタリングを行って，図 1 の結果と比較を行った．

図 4 は Bag-of-Foods でユーザの食習慣を表現し，K-means でクラスタリングを行った結果である．図 1 におけるクラスタをユーザごとの栄養価属性と呼ぶことにして，図 4 の各クラスタに占める栄養価属性の割合を求めたところ，平均して 83.4%のユーザがクラスタごとには同じ栄養価属性を持つことが確認でき，基本的には平均的な栄養の偏りが表現できることが分かった．

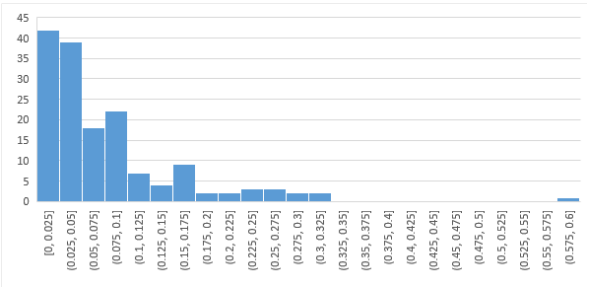

図 5: 同一ユーザで異なる時期の Bag-of-Foods 特徴ベクトルの距離の分布

また，栄養価属性の分布はクラスタの性質によって異なることが確認できた．平均的に量が多いあるいは少なかったり，全体的に栄養に偏りがあるクラスタはほとんどが同じ栄養価属性のユーザで構成されており，全員が同じ属性であるケースも少なくない．一方で間食の割合が多いなど，食事の意味的な分布に特徴があるクラスタは栄養価属性が統一されていなかった．手法から自明ではあるが，提案特徴量が一食ごとの細かな違いよりも食事属性の分布を重視する傾向が現れている．

表 2: 食習慣が異なるといえる食事記録の例 - 特徴量間の距離: 1.005

ユーザ A	ユーザ B
卵サンド	トマトサラダ, バタートースト, ブラックコーヒー, 三色ナムル
おはぎ	鶏のから揚げ, けんちんうどん, りんご
ベイクトチーズケーキ, カスタードプリン, 野菜サンド, ブラックコーヒー	味付けメンマ, 鶏のから揚げ, ほうれん草のおひたし, ご飯
パンケーキ, ブラックコーヒー	ほうれん草のソテー, 豚の角煮, りんご, ご飯
かぼちゃの田舎煮, 味噌汁, 切干し大根の炒め煮, ししゃも, 雑穀ご飯	パンケーキ

ところで，栄養価属性の中でも図 1 におけるクラスタ 0 とクラスタ 4・5 は特に食習慣がかけ離れており，食習慣ベクトル上でもこれらのユーザは離れているべきである．これらの栄養価属性を持つユーザ間での距離を調べたところ，すべてのケースにおいて距離が 1.0 以上離れていることが分かった．表 2 の例に示す通り，食習慣ベクトルでは距離が 1.0 以上あれば食習慣が明らかに異なっていると言え，提案手法が異なる食習慣を異なるように表現できていることが確認できる．

4.2. 食習慣の同一性

食習慣は同一人物のものでも時間的に変化する．とはいえ，その変動は全体で見れば比較的小さいはずである．そこでユーザごとの食事記録を 2 つに分割して，それぞれに独立して特徴量を作り，同一人物・異なる時期の食習慣が特徴量空間上でどの程度離れているか調べた．

表 3: 食習慣が近いといえる食事記録の例 - 特徴量間の距離: 0.248

ユーザ A	ユーザ B
バナナ	ドライマンゴー
グリーンサラダ, おにぎり	コーンフレーク, バナナ, ヨーグルト
にゅうめん, うぐいす豆の甘煮, 切干し大根の煮物	ちくわの天ぷら, トマトスパゲッティ, ご飯, なし, 鶏のから揚げ
野菜炒め, まぐろの刺し身	肉野菜炒め, 焼きとり, ツナサラダ, 味噌汁, 麦茶
バナナ, ぶどう	ヨーグルト, バナナ, ツナサラダ, コーンフレーク
味噌汁, 茶碗蒸し, 焼き魚, 麦とろ飯, れんこんのきんぴら	卵雑炊, なし, 鶏の照り焼き, えびマカロニグラタン
チョコレートクッキー	キャラメル

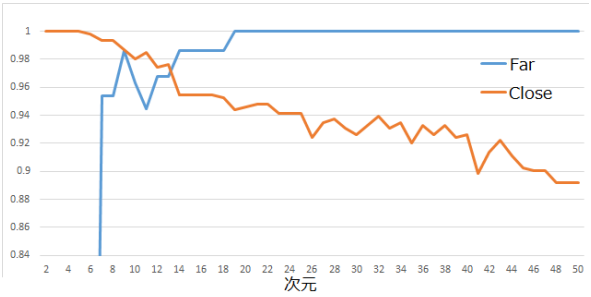

図 6: 特徴量次元数に対する離れるべき特徴同士が遠い割合 (Far) と同一ユーザの特徴同士が近い割合 (Close) の推移

まず，全ユーザ 157 人 (156 人の他ユーザ+異なる時期の自分自身) のベクトルのうち，自身の別の時期のベクトルが自身のベクトルに最も近くなるケースは 76.6% となった．ただし，残り 23.4%のケースについて，自身より近いとされたユーザの実際の食事記録を確認してみると，食習慣の類似度としてはより近いように思われた．そこで，同一人物・異なる時期の食習慣同士の距離が「類似の食習慣であると言えそうな距離の閾値」を超えない割合を求めた．今回，この閾値は 0.25 に定めた．表 3 に 0.25 程度離れている食習慣の実際の食事記録の一部を示す．この例では，間食・軽食・通常の食事のバランスが似通っているほか，各食事の栄養バランスもある程度似ていることが確認でき，0.25 の距離は類似の食習慣といえそうである．さて，同一ユーザについて異なる時期の二つの食習慣ベクトルの距離をそれぞれ求めると，その分布は図 5 のようになった．距離が 0.25 以内であるユーザは 94.8%となった．提案手法において，同一ユーザの時間的な食習慣の揺らぎはほとんど吸収できていることが分かる．

4.3. 特徴量次元数の選定

4.1 節と 4.2 節にて，定性的ではあるが離れるべき・近づくべき特徴量がそう振舞っている割合を示す指標が得られた．これをもとに，提案特徴量の次元数を変更し，これらの指標がどう変化するか検証した．この結果を図 6 に示す．2 つの指標はおおむねトレードオフの関係にあるが，今回の指標では近さの保証がない近付き方に基づく指標よりも離れ方に基づく指標のほうがより信頼できる．以上より離れ方の指標を重視すると，こちらが最初に飽和する 20 次元が最も良い特徴量になると結論付けられる．

5. まとめと今後の課題

本稿では大規模食事記録を使用して，摂取栄養価をもとにしたユーザのクラスタリングや食習慣のベクトル化の手法について論じた．まず，一食当たりの平均摂取栄養価をもとにユーザの分類を行ったところ，小食の群，過食の群，野菜偏重の群といった解釈可能なクラスタが得られ，食習慣の分析においては摂取栄養価に基づいたアプローチが妥当なものであることを確認した．次に，同様のクラスタリングで得た一食ごとの食事属性をもとにユーザの食習慣をベクトル化する Bag-of-Foods を提案し，その性質について論じた．

今後の課題として，まず時系列情報を残せるベクトル化手法を取り入れることが挙げられる．本手法では時間情報がつぶれてしまっており，密な記録を付けるユーザと疎な記録を付けるユーザの区別がついていないほか，朝食・昼食・夕食といった食事の時間帯を考慮できていない．また，現状では食習慣ベクトルの定量的な評価ができていないため，そのための評価手法も必要である．§

参考文献

[1] K. Aizawa and M. Ogawa. Foodlog: Multimedia tool for healthcare applications. *IEEE MultiMedia*, Vol. 22, No. 2, pp. 4–8, Apr 2015.

[2] E. Kesse-Guyot, V. A. Andreeva, C. Lassale, S. Hercberg, and P. Galan. Clustering of midlife lifestyle behaviors and subsequent cognitive function: A longitudinal study. *American Journal of Public Health*, Vol. 104, No. 11, pp. e170–e177, 2014. PMID: 25211733.

[3] Maria C. Patino-Alonso, Jos I. Recio-Rodrguez, and et al. Clustering of lifestyle characteristics and their association with cardio-metabolic health: the lifestyles and endothelial dysfunction (evident) study. *British Journal of Nutrition*, Vol. 114, No. 6, p. 943951, 2015.

[4] National Cancer Institute. Diet history questionnaire ii. `https://epi.grants.cancer.gov/dhq2/`.

[5] 椿田晃大, 天野宗佑, 相澤清晴, 小川誠. 食事記録データからの個人の食傾向の予測. 映像情報メディア学会技術報告 = ITE technical report, Vol. 41, No. 16, pp. 55–60, 2017.

[6] 「日本人の食事摂取基準（2015 年版）策定検討会」報告書. Mar 2014.

§本研究の一部は，JST CREST (JPMJCR1686)，科研費 (18H03254) の支援を受けた．