D-005

行動時間帯に偏りのある長時間エピソード抽出における発生区間の範囲による探索候補の枝刈りの提案

Method of pruning candidates by considering amount of appearing time for mining long-duration episodes with biased occurrence time-range

安井 壱陽 † 新谷 隆彦 † 大森 匡 † 藤田 秀之 †
KazuhiYasui Takahiko Shintani Tadashi Ohmori Hideyuki Hujita

1. 背景と目的

近年のスマートフォンなどの小型端末の普及により, 人の生活に関するデータであるライフログを収集することが容易になった. スマートフォンの操作履歴などのライフログから日常的な利用行動を抽出するモバイルマイニング [1] などのライフログの活用に関する研究が進められている. 我々は, リストバンド型加速度センサから取得した腕の動きに関するライフログからエピソードマイニングを用いて日常的な行動に相当する長時間エピソードを抽出し, 生活特性を検出する研究 [2] がある. 長時間エピソードマイニングでは長い時間費やした行動に相当する長時間エピソードを抽出する. 報告者は行動した時刻に着目し, 特定の時間帯によく行われた行動に相当するエピソードとして, 行動時間帯の偏りを考慮した長時間エピソードを検討してきた [3]. しかし, 行動時間帯の偏りを考慮した長時間エピソードは 1 日の中での限られた時間帯でよく行われた行動に対応したエピソードとなるため, 行動時間帯が偏った長時間エピソードの抽出では従来の長時間エピソードマイニングよりも閾値を低く設定しなければならない. そのため, 探索候補数が多くなり処理負荷が高くなることが問題であった.

そこで本研究では, 探索負荷を低減するための探索候補の枝刈り手法を提案する. 探索候補の発生区間と時間帯が重なり得る最大の区間を計算し, 探索候補を低減する. 報告者が収集したライフログデータを用いた評価実験より, 処理負荷を低減可能なことを確認する.

2. 長時間エピソード

長時間エピソード [2] の定義を示す. 処理対象データをシーケンスデータ $S = \langle e_1, e_2, ..., e_n \rangle$ とし, これはイベント e を開始日時順に並べたリストである. イベント $e = (\epsilon, t_s, t_e)$ はイベントタイプ ϵ と開始日時 t_s, 終了日時 t_e の組からなり, ϵ が t_s から t_e まで行われたことを表す. ここで, シーケンスデータ上のイベント (ϵ, t_s, t_e) に対し, $t'_s < t_e$ かつ $t_s < t'_e$ となるイベント (ϵ', t'_s, t'_e) は存在しない.

エピソード $\alpha = \langle a_1\text{-}a_2\text{-}\cdots\text{-}a_k \rangle$ はイベントタイプのリストである. a_i は $a_i \in E(1 \leq i \leq k)$ を満たす. エピソード α を構成する各イベントタイプを含むイベントがシーケンスデータ S に同一の順序で現れた時, エピソードがシーケンスデータに発生したと表現する. つまり, 長さ k のエピソード $\alpha = \langle a_1\text{-}a_2\text{-}\cdots\text{-}a_k \rangle$ は, シーケンスデータ $S = \langle (\epsilon_1, t_{s_1}, t_{e_1}), (\epsilon_2, t_{s_2}, t_{e_2}), \cdots, (\epsilon_n, t_{s_n}, t_{e_n}) \rangle$ に対し, $a_i = \epsilon_{j_i}$ $(1 \leq j_1 < j_2 < \cdots j_k \leq n,\ 1 \leq i \leq k)$ を満たすとき, α が S に発生したとする. エピソード α が発生した区間 $(\epsilon_{j_1}, t_{s_{j_1}}, t_{e_{j_1}}), (\epsilon_{j_1+1}, t_{s_{j_1+1}}, t_{e_{j_1+1}}), \cdots, (\epsilon_{j_k}, t_{s_{j_k}}, t_{e_{j_k}})$ をオカレンスと呼び, $[t_{s_{j_1}}, t_{e_{j_k}})$ と表す. オカレンスの継続時間は $t_d = t_{e_{j_k}} - t_{s_{j_1}}$ である. オカレンスには制約条件として, 最大継続時間 $maxspan$ と最大ギャップ $maxgap$ を用いる. 最大継続時間により継続時間が極端に長いオカレンスを除外する. 最大ギャップによりイベント間の時間間隔が長いオカレンスを除外する.

エピソード α は頻度と総継続時間を評価値として持つ. α の極小非重複オカレンス [4] の数が頻度であり, α の極小非重複オカレンスの継続時間の総和が総継続時間である. 極小オカレンスとは開始日時から終了日時の間に他のオカレンスを含まないオカレンスであり, 極小オカレンスの集合を $MO(\alpha)$ と表す. つまり, エピソード α について, オカレンス $occ = [t_s, t_e)$ が $t_s \leq t'_s$ かつ $t'_e \leq t_e$ を満たすオカレンス $[t'_s, t'_e)]$ が存在しない時 occ は極小オカレンスとなる. 非重複オカレンスは, オカレンスが互いに重複していないことを表す. つまり, エピソード α について, オカレンス $occ = [t_s, t_e)$ が $t_e < t'_s$ かつ $t'_s < t_e$ を満たすオカレンス $[t'_s, t'_e)]$ が存在しない時 occ は非重複オカレンスとなる. 極小かつ非重複なオカレンスを極小非重複オカレンスと呼び, その集合を $MANO(\alpha)$ と表す.

3. 行動時間帯に偏りのあるエピソード

行動時間帯に偏りのあるエピソードの定義を示す. 時刻 x と y の間を時間帯と呼び, 日々の x から y までの時間帯 $r = [t_s, t_e)$ の集合を行動時間帯を T_r とする. 時間帯の幅は $r_e - r_s$ 分である. 本研究では, 時間帯 r の幅, 最大オカレンス継続時間が 720 分未満とする.

エピソード α のオカレンス $occ(\alpha) = [t_s, t_e)$ がある $r = [r_s, r_e)$ に対して $t_s < r_e$ かつ $r_s < t_e$ を満たす時, 時間帯 $[t_s, t_e)$ でオカレンスが発生したとし, オカレンスが時間帯と重なった時間を時間帯継続時間 r_d とする. 総時間帯継続時間は α のすべての極小非重複オカレンスの r_d の総和である. r_d は $r_e - r_s$ と計算する. ここで, $t_s < r_s$ の時 $t_s = r_s$, $r_e < t_e$ の時 $t_e = r_e$ する. $t_s < r_e$ かつ $r_s < t_e$ なる時間帯が r に存在しない場合, $r_d = 0$ とする.

エピソード α は総時間帯継続時間, 時間帯偏り度を評価値として持つ. α の極小非重複オカレンスの時間帯継続時間の総和が総時間帯継続時間である. エピソー

†電気通信大学大学院情報理工学研究科 Graduate School of Informatics and Engineering, The University of Electro- Communications

ドが所定の時間帯に偏っているかを評価するための評価値を時間帯偏り度 T_b と呼び, 以下の式で定義される.

$$T_b = \frac{\sum_{o \in MANO(\alpha)} r_d(o)}{\sum_{o \in MANO(\alpha)} t_d(o)}$$

$r_d(o)$ は α の極小非重複オカレンス o の時間帯継続時間, $t_d(o)$ は α の極小非重複オカレンス o の継続時間を表す. 時間帯偏り度は値が高い程その時間帯に大きく偏ってエピソードが発生したことを示す.

行動時間帯に偏りのある長時間エピソード抽出問題は, ユーザが指定した時刻 t_s から t_e までの日々の時間帯の集合 T_r に対して, 最小時間帯偏り度 $mintb$, 最小総時間帯総継続時間 $mintrdur$ を満たすエピソードをすべて抽出することである. また, オカレンスの制約条件として最大オカレンス継続時間 $maxspan$, 最大ギャップ $maxgap$ も設定する.

次に, 行動時間帯に偏りのある長時間エピソード抽出の手順を説明する. エピソード α の末尾にに 1 つのイベントタイプ e を追加してエピソードを成長させ, このエピソードのオカレンスを $MO(\alpha)$ と $MO(e)$ から作成し, 最小総時間帯継続時間および最小時間帯偏りを満たすかどうかを調べることによって行動時間帯に偏りのある長時間エピソードを抽出する. Algorithm1 にアルゴリズムを示す.

Algorithm 1: Extracttbepisodes

```
1  S を 1 回スキャンし, すべてのイベントタイプの
     極小オカレンスを生成
2  E := lowfreq を満たすすべてのイベントタイプ
3  T_r :=S の開始日時から終了日時までの時間帯の
     集合
4  foreach e ∈ E do
5      rdur(e) = Calctrdur(e)
6      if trdur(e) ≥ mintrdur then
7          tdur(e) = Σ_{[t_s,t_e)∈MAMO(e)} (t_e − t_s)
8          if trdur(e)/tdur(e) ≤ mintb then
9              output e
10         end
11     end
12     foreach m ∈ E do
13         Extend(e, m)
14     end
15 end
```

始めにシーケンスデータを 1 回スキャンし, すべてのイベントタイプの極小オカレンスを生成する. その後, 頻度が $lowfreq = \frac{mintdur}{maxspan}$ 以上のイベントタイプの集合を E を生成する. そして, イベントタイプ $e \in E$ が最小総時間帯継続時間と最小時間帯偏り度を満たすかを判定し, 満たす場合は行動時間帯に偏りのある長時間エピソードとして出力する. その後, $m \in E$ に対して $Extend(e, m)$ を呼び出し, エピソードを成長させる.

$Extend$ はエピソード α の末尾にイベントタイプ e を追加して成長させたエピソード β を調べる関数である. Algorithm2 に関数 Extend を示す. α の末尾に e

Algorithm 2: Extend(α, e)

```
1  β := α の後に e が発生したエピソード
2  MO(β) = GenerateMo(MO(α), MO(e))
3  MAMO(β) = GenerateMano(MO(β))
4  trdur(β) = Calctrdur(MANO(β))
5  if trdur(β) ≥ mintrdur then
6      tdur(β) = Σ_{[t_s,t_e)∈MANO(β)} (t_e − t_s)
7      if trdur(β)/tdur(β) ≥ mintb then
8          output β
9      end
10 end
11 if freq(β) ≥ lowfreq then
12     foreach m ∈ E do
13         Extend(β, m)
14     end
15 end
```

を接続することで新たなエピソード β を生成し, 関数 GenerateMO(Algorithm3) で β の極小オカレンスを, 関数 GenerateMAMO(Algorithm4) で極小非重複オカレンスを生成する. そして, β の総時間帯継続時間を Calctrdur で計算し, 最小総時間帯継続時間以上であれば総継続時間及び時間帯偏り度を計算する. Calctrdur はエピソードの極小非重複オカレンスの集合から総時間帯継続時間を計算する関数である. 最小時間帯偏り度を満たしていれば β を出力する. 長時間エピソードは頻度ではなく総継続時間で評価するが, 総継続時間は Apriori の性質を満たさないため, 従来の探索候補の枝刈りはできない. しかし, 最小総継続時間を満たしかつ最も頻度が少ない場合は, すべてのオカレンスの継続時間が $maxspan$ となるときである. 最小総継続時間を満たさないエピソードは最小総時間帯継続時間も満たさないため, 下限頻度 $lowfreq = mintdur/maxspan$ を満たさないエピソードの探索を除外することができる. これを利用して, Algorithm2 では頻度が $lowfreq$ 以上の $m \in E$ に対して $Extend(\beta, m)$ を呼び出し, 探索を継続する.

4. 提案手法

行動時間帯に偏りのあるエピソードは特定の時間帯における継続時間で評価するため, 従来の長時間エピソードの最小総継続時間に対応する閾値である最小総時間帯継続時間を低く設定する必要がある. つまり, 下限頻度 $lowfreq$ が小さくなるため Algorithm1 では探索候補数が多くなってしまう. そこで, 本研究では探索候補のオカレンスを作成する前に, その探索候補がとり得る総時間帯継続時間の最大値を求めることによって, 探索候補を枝刈りする.

エピソード α にイベントタイプ e を追加して成長させた探索候補の総時間帯継続時間の最大値は, α の極小オカレンス $MO(\alpha)$ と $maxspan$ から求めることができ

Algorithm 3: GenerateMO(α, e)

$MO(\beta) = \phi$
foreach $[t'_s, t'_e) \in MO(e)$ **do**
  $t_{e_j} \leq t'_s$ かつ $t'_s - t_{e_j} \leq maxgap$ かつ $t'_e - t_s \leq maxspan$ かつ $t_{e_{j+1}} > t'_s$ となるような $[t_{s_j}, t_{e_j}) \in MO(\beta)$ を見つける
  if $[t_{s_j}, t_{e_j})$ が存在 **then**
    $MO(\beta)$ に $[t_{s_j}, t'_e)$ を加える
  end
end
output $MO(\beta)$

Algorithm 4: GenerateMANO($MO(\beta)$)

$i = 0$
$j = 1$
$MAMO(\beta) = \phi$
$MAMO(\beta)$ に $[t_{s_0}, t_{e_0})$ を加える
while $j < |MO(\beta)|$ **do**
  $[t_{s_i}, t_{e_i}) \in MO(\beta)$ かつ $[t_{s_j}, t_{e_j}) \in MO(\beta)$ かつ $t_{e_i} \leq t_{s_j}$ となる最初の j を見つける
  if $t_{e_i} \leq t_{s_j}$ **then**
    $MANO(\beta)$ に $[t_{s_j}, t_{e_j})$ を加える
    $i = j$
    $j = j + 1$
  end
end
return $MANO(\beta)$

る．探索候補の個々のオカレンスがとり得る時間帯継続時間の最大値は 3 通りに分けられる．オカレンス $occ = [t_s, t_e)$ について，その開始日時から $maxspan$ 分後の日時を終了日時としたオカレンスを $occ' = [t_s, t'_e)$ とする．occ' のオカレンス開始日時が時間帯内に存在する場合，つまり $r_s \leq t'_s < r_e$ となる $[r_s, r_e) \in T_r$ が存在する場合，時間帯継続時間の上界は $min\{r_e - t'_s, maxspan\}$ である．occ' のオカレンス開始日時が時間帯外かつオカレンスが時間帯と重なる場合，つまり $t'_s < r_s$ かつ $r_s < t'_e$ となる $[r_s, r_e) \in T_r$ が存在する場合，時間帯継続時間の上界は $min\{r_e - r_s, (t'_e) - r_s\}$ である．occ' が時間帯と重なっていない場合，つまりすべての $[r_s, r_e) \in T_r$ に対して $t'_s < r_e$ または $r_s < t_e$ のどちらかしか満たさない場合，時間帯継続時間の上界は 0 である．例として，時間帯を 12 時から 18 時まで，$maxspan$ を 420 分としたときのオカレンスの取り得る時間帯継続時間の範囲を図 1 に示す．横軸は時刻，縦軸は日付を表している．オカレンスの時刻が 12 時から 18 時の間の部分が時間帯継続時間のとり得る範囲となる．

探索候補の総時間帯継続時間が最小時間帯総継続時間を満たすためには，探索候補のとり得る総時間帯継続時間の最大値が最小時間帯総継続時間を満たさなければならない．探索候補のオカレンスがとり得る継続時間の最大値はすべて $maxspan$ であるが，とり得る時間帯継続時間の最大値はオカレンスごとに異なり，$maxspan$ より小さくなることが期待される．そこで提案手法では下限頻度 $lowfreq$ ではなく，とり得る総時間帯継続時間の最大値によって探索候補の枝刈りを行う．

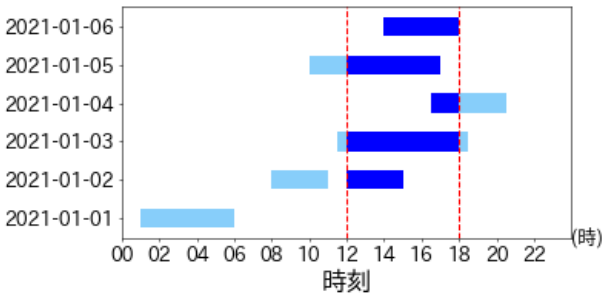

図 1: 時間帯継続時間のとり得る範囲の例

提案手法では，探索候補を成長させる前に取り得る総時間帯継続時間による枝刈りを行う．Algorithm1 では，11 行目と 12 行目の間に $upb(e) = CalcUpb(MO(e))$ を追加し，$upb(e) \geq mintrdur$ を満たす場合にのみ 12 行目から 14 行目の処理を行う．また，Algorithm2 の 10 行目と 11 行目の間に $upb(\beta) = CalcUpb(MO(\beta))$ を追加し，$upb(\beta) \geq mintrdur$ を満たす場合にのみ 11 行目から 15 行目の処理を行う．ここで，CalcUpb は探索候補のエピソードがとり得る総時間帯継続時間の最大値を計算する関数である．

5. 評価実験

提案手法の処理負荷の低減の評価実験を行った．従来の長時間エピソードマイニングの下限頻度 $lowfreq$ による枝刈りをベースラインとし，提案手法のベースラインに対する処理負荷を低減を実験により評価した．本実験では，処理対象のデータとして報告者が収集した 2019 年 04 月 11 日から 2020 年 2 月 05 日までの 301 日分の生活様式が大きく変わる前までの運動状態データ用いた．運動状態データとは，リストバンド型センサで取得したいつからいつまでどんな運動状態が継続したかを表すデータである．運動状態には，”静止”，”安静”，”デスクワーク”，”軽作業”，”作業”，”歩行”，”ジョギング”，”運動”，”非装着”がある．運動状態データは長時間エピソードマイニングにおけるイベント，運動状態はイベントタイプに相当するため，運動状態データを開始時刻順に並べたリストはシーケンスデータとなる．また，本実験において最小時間帯偏り度を 0.30，最大オカレンス継続時間を 360 分，最大ギャップを 60 分とした．時間帯には幅が 6 時間となる時間帯を 4 通り，4 時間となる時間帯を 2 通り用いた．最小総時間帯継続時間は，時間帯の幅が 6 時間の場合 1 週間当たり 60 分，4 時間の場合 1 週間当たり 40 分とした．処理時間は 100 回の計測の平均値としている．

データの期間を変化させた場合の処理時間を図 2 と図 3 に示す．ここで，データの期間は，開始日を 2019 年 04 月 11 日とし，終了日を 2019 年 08 月 05 日から 2020 年 2 月 05 日まで 1 か月ずつ延ばした場合とした．図 2 はベースラインアルゴリズム，図 3 は提案手法の処理

時間である．横軸は 2019 年 04 月 11 日から 2019 年 08 月 05 日までの運動状態データの番号を 1 として終了日が早い順に割り振ったデータの番号を表し，縦軸は処理時間を表している．

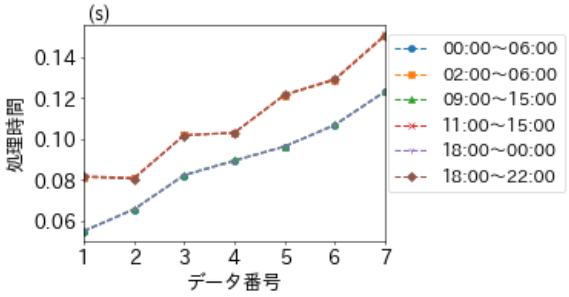

図 2: ベースラインの処理時間

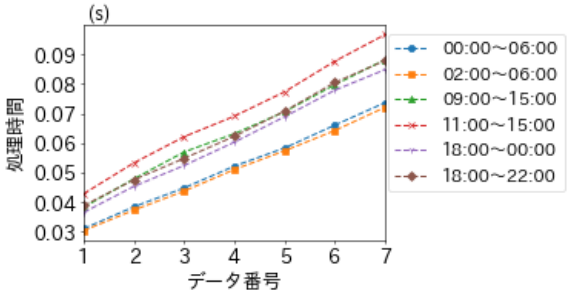

図 3: 提案手法の処理時間

さらに，ベースラインに対して提案手法の処理時間がどの程度減少したかを図 4 に示す．横軸は図 2 と同じくデータ番号を，縦軸は減少率を表している．減少率はベースラインの処理時間を t，提案手法の処理時間を t' とし $-\frac{(t'-t)}{t}$ で計算され，値が高いほど大きく処理時間が減少していることを示す．

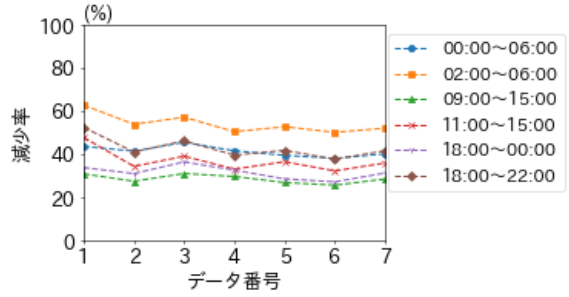

図 4: ベースラインに対する処理時間の減少率

図 2 より，ベースラインアルゴリズムにおいて時間帯の幅が等しい場合，処理時間に差は見られないが，図 3 より，提案手法では時間帯の幅が等しい場合も処理時間に差が見られる．2 つのグラフを比較すると，すべてのデータにおいて提案手法のほうが処理時間が短くなっている．ベースラインと提案手法の処理時間の差は，探索したエピソード数に依存していると考えられる．そこで，探索したエピソード数を図 5 にベースライン，図 6 に提案手法に関して示す．横軸は図 2 と同じくデータ番号を，縦軸は探索したエピソード数，つまり枝刈りされずに残った候補の数を表している．

図 5 より，ベースラインアルゴリズムでは時間帯が異なっていても探索エピソード数が等しい場合がある．これは，下限頻度が最小総時間帯継続時間と最大オカレンス継続時間に依存するためである．本実験では，時間帯の幅ごとに最小総時間帯継続時間を変化させていたため，時間帯の幅が等しい場合に探索エピソード数が等しくなった．処理時間が等しくなっていた原因も同様である．図 5 と図 6 を比較すると，探索エピソード数が大幅に減少していることが分かる．さらに，ベースラインアルゴリズムは提案手法に比べ，候補数がデータの期間によって上下している．これは下限頻度が極端に低くなり，細かいデータの傾向の変化で候補数が変化するためであると考えられる．

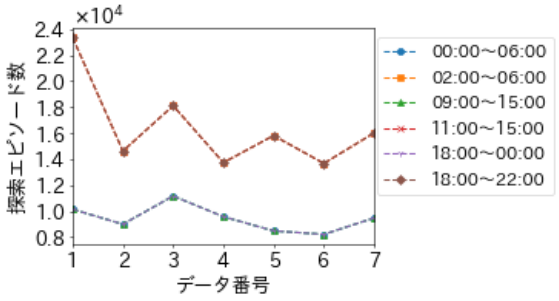

図 5: ベースラインの探索エピソード数

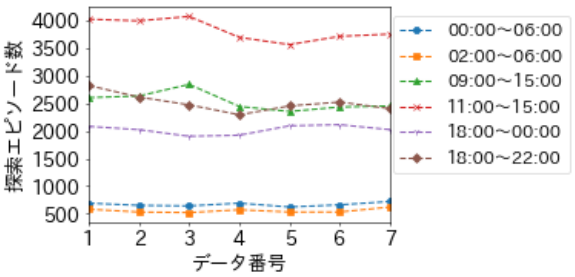

図 6: 提案手法の探索エピソード数

6. 終わりに

本研究では，行動時間帯に偏りのある長時間エピソードを抽出する際に，探索候補の増大により処理負荷が高くなる問題点の解決を行った．とり得る時間帯継続時間の最大値が，最大オカレンス継続時間より短くなることに着目した探索候補枝刈り手法を提案した．報告者のデータを用いた評価実験において，ベースラインとした手法より有効であることを示した．

謝辞

本研究は JST, CREST の支援を受けたものである．

参考文献

[1] R. Rawassizadeh, E. Momeni, C. Dobbins, J. Gharibshah, M. Pazzani, *Scalable Daily Human Behavioral Pattern Mining from Multivariate Temporal Data*, IEEE TKDE, 2016.

[2] T. Shintani, T. Ohmori, H. Fujita, *Method for Comparing Long-term Daily life using Long-duration episodes*, EDBT/ICDT Workshops 2019.

[3] 安井壱陽, 新谷隆彦, 大森匡, 藤田秀之, ”継続時間を考慮したエピソードマイニングにおける行動時間帯の偏りに関する一考察”, 情報処理学科会第 82 回全国大会, 4N-04, 2020

[4] H. zhu, P. Wang, X. He, Y. Li, W. Wang, B. Shi, *Efficient Episode Mining with Minimal and Non-overlapping Occurrences*, IEEE ICDM, 2010.