D-020

長時間エピソードマイニングにおける 冗長な重複の回避によるオカレンス作成処理の削減手法

Method for Reducing Occurrence Creation Processes by Avoiding Redundant Overlaps in Long Duration Episode Mining

橋本 一輝 † 新谷 隆彦 † 大森 匡 † 藤田 秀之 †
Kazuki Hashimoto Takahiko Shintani Tadashi Ohmori Hideyuki Fujita

1. 背景と目的

近年のスマートウォッチ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データの収集が容易になった. 我々は, ライフログデータを一つの長いイベントシーケンスとみなし, エピソードマイニングを用いて, 長時間行った運動状態のパターンを抽出する研究を行っている [1]. 長時間エピソードマイニングでは, エピソードが出現した区間であるオカレンスを作成し, エピソードが出現した総継続時間を計算することにより, 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. 長時間エピソードマイニングの評価値である総継続時間は Apriori の性質を満たさず, 総継続時間を用いて枝刈りをすることができない. そのため, 長時間エピソードマイニングでは, Episode-Weighted Utilization(EWU)[2] と呼ばれる, エピソードが最大まで成長したときの評価値の上限を利用することで候補エピソードの枝刈りを行う.

単純な EWU は, 重複した区間を複数回足し合わせる計算により, 冗長な総継続時間の上限値になる場合がある. 本稿では, 重複を考慮した計算により導出される上限値を作成し, これを導入することで長時間エピソードマイニングの探索処理の負荷を低減させることを検討する.

2. 長時間エピソードマイニング

シーケンスデータ $D = \langle d_1, ..., d_n \rangle$ は, 運動状態データを開始日時の順に並べたリストである. 運動状態データ $d_i = (m_j, t_{s_i}, t_{e_i})$ は, 開始日時 t_{s_i} から終了日時 t_{e_i} までの間, 運動状態が $m_j \in M$(M は運動状態の集合) であったことを意味する.

エピソード $\alpha = \langle m_1, \ldots, m_k \rangle$ は運動状態のリストであり, $m_i \in M (1 \leq i \leq k)$ である. シーケンスデータ D の中にエピソード α を構成するそれぞれの運動状態 m_j が同じ順序で表れるとき, α が D の中に出現すると呼ぶ. このときの α の出現区間をオカレンスと呼び, $occ(\alpha)$ と表す. $occ(\alpha)$ は, m_1 の運動状態データの開始日時と m_k の運動状態データの終了日時を用いて, $[t_{s_{m1}}, t_{e_{mk}})$ と表す. $occ(\alpha)$ の継続時間は, $t_{e_{mk}} - t_{s_{m1}}$ となる. 冗長なオカレンスを除くため, 一つのオカレンスの継続時間の上限値 $maxspan$ と, オカレンスの中の運動状態データの間隔の上限値 $maxgap$ が定義され, これらの制約を満たすようなエピソード α のオカレンスの集合を $OCC(\alpha)$ と表す. $OCC(\alpha)$ のオカレンスについて, $occ(\alpha)$ が別の α のオカレンスを含まないとき, 極小であると呼び, α の極小なオカレンスの集合を $MO(\alpha)$ と表す. また, 2 つのオカレンスの区間が重複していないとき, その 2 つのオカレンスの関係を非重複と呼び, 極小かつ非重複な α のオカレンスの集合を $MANO(\alpha)$ と表す.

エピソードは頻度と総継続時間を評価値として持つ. エピソード α の頻度は $freq(\alpha) = |MANO(\alpha)|$ と定義される. また, エピソード α の総継続時間は $tdur(\alpha) = \sum_{[t_s,t_e] \in MANO(\alpha)} (t_e - t_s)$ と定義される. エピソードの総継続時間がユーザー指定の総継続時間の最小値 $mintdur$ 以上であるとき, そのエピソードを長時間エピソードと呼ぶ. 長時間エピソードマイニングは, シーケンスデータ D, 総継続時間の閾値 $mintdur$, 一つのオカレンスの継続時間の上限値 $maxspan$, オカレンスの中の運動状態の間隔の上限値 $maxgap$ が与えられたときに, 全ての長時間エピソードを抽出する問題である.

3. 従来手法

エピソードを成長させたときに総継続時間が長くなる場合があるため, 長時間エピソードマイニングでは単純な Apriori の方法で枝刈りをすることができない. しかし, エピソードが長時間エピソードである場合, 全てのオカレンスの継続時間が $maxspan$ となるときに頻度が最小になるため, 下限頻度 $lowfreq = \lceil \frac{mintdur}{maxspan} \rceil$ を設定し, 下限頻度を満たさない候補エピソードを枝刈りすることができる [1].

長時間エピソードマイニングのアルゴリズムは, 初めにデータベースを 1 度だけ読み取り, 各運動状態について 長さ 1 のエピソードの極小なオカレンスの集合を作成し, 下限頻度を満たす運動状態を残す. その後, 既に作成したエピソード α の末尾に, 長さ 1 のエピソード m を結合することで探索候補となるエピソード $\beta = \langle \alpha, m \rangle$ を深さ優先で作成していく. 作成したエピソードについて $MO(\beta)$, $MANO(\beta)$, 総継続時間を計算し, 長時間エピソードを抽出する. $ExtractLongDurationEpisodes$ に長さ 1 の極小なオカレンスを作成する手順を示し, $ExtendEpisode$ に長時間エピソードの探索の手順を示す.

$ExtractLongDurationEpisodes(D, minfreq, mintdur, maxspan, maxgap)$

1. min_tdur と $maxspan$ から頻度の下限として, $lowfreq = \lceil \frac{mintdur}{maxspan} \rceil$ を計算する.

†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

2. データベース D を 1 度だけ読み取り, 長さ 1 のエピソードについて, 極小なオカレンスを作成し, その要素数が $lowfreq$ を満たすものを FI として保持する. その後, 全てのエピソードの組 $(h,m) \in FI$ について $ExtendEpisode(h,m)$ を呼び出す.

$ExtendEpisode(\alpha, m)$

1. 長さ k のエピソード α の末尾に長さ 1 のエピソード m を連結することで, 長さ $k+1$ のエピソード β を作成する.
2. $MO(\alpha)$ と $MO(m)$ から $MO(\beta)$ を計算する.
3. $MO(\beta)$ から $MANO(\beta)$ を計算する. $|MANO(\beta)| \geq lowfreq$ であれば, 全ての長さ 1 のエピソード m' について, $ExtendEpisode(\beta, m')$ を呼び出す. その後, $MANO(\beta)$ から, 総継続時間を計算し, その値が $mintdur$ 以上であれば, 長時間エピソードとして出力する.

下限頻度 $lowfreq$ は, ユーティリティエピソードマイニングにおける Episode-Weighted Utilization(EWU)[2] を計算することと等しい. 長時間エピソードマイニングにおけるエピソード α の EWU の値は, $EWU(\alpha) = maxspan \times |MANO(\alpha)|$ によって計算される. 従来手法における枝刈りでは, 全てのオカレンスの継続時間が $maxspan$ まで延びる場合の総継続時間を計算するが, この計算の途中で重複した区間を複数回足し合わせ, 重複のある冗長なオカレンスから, 緩い上限値が計算される場合がある.

4. 提案手法

長時間エピソードマイニングは, 極小かつ非重複なオカレンスから総継続時間の計算を行うが, 従来の EWU の計算では非重複の条件を考慮せず, 重複のある部分を複数回足し合わせる場合がある. ここで, オカレンス同士の位置関係を確認しながら, 重複する区間を足し合わせずに, 総継続時間の上限値を計算すれば, 長時間エピソードマイニングの探索空間を削減できると考えた.

本稿では新しい閾値として, 非重複を考慮する EWU ($NOEWU$) を提案する. エピソード α について, $MO(\alpha)$=$\{[t_{s_1}, t_{e_1}], [t_{s_2}, t_{e_2}], ..., [t_{s_n}, t_{e_n}], [t_{s_{n+1}}, t_{e_{n+1}}], ..., [t_{s_k}, t_{e_k}]\}$ とするとき, α の $NOEWU$ を次のように定義する.

$$NOEWU(\alpha) = \sum_{n=1}^{k-1} min\{(t_{s_{n+1}} - t_{s_n}), maxspan\} + maxspan$$

$NOEWU$ は, α のそれぞれの極小なオカレンスの継続時間の上限が, 次のオカレンスの開始日時までの時間と $maxspan$ の小さい方であることを利用して導出する.

従来手法は極小かつ非重複なオカレンスから EWU を計算するが, 提案手法では極小なオカレンスの集合から $NOEWU$ を計算する. エピソード α, α を成長させたエピソード β について, 常に $|MO(\alpha)| \geq |MO(\beta)|$ であり, $MO(\alpha)$ の開始日時の集合は $MO(\beta)$ の開始日時の集合を含む. オカレンスの開始日時の集合の要素が単に減少するとき $NOEWU$ の値は単調に減少するため, 極小なオカレンスの集合から計算される $NOEWU$ によって枝刈りをすることができる.

長時間エピソードマイニングの ExtendEpisode で作成される極小なオカレンスに対して, $NOEWU$ を計算することにより, その値が閾値を満たさなかった場合の処理を省略することができる. 提案手法では $ExtendEpisode$ の手順 2. を以下のように変更する.

2'. $MO(\alpha)$ と $MO(m)$ から $MO(\beta)$ を計算し, 保持する. $MO(\beta)$ から $NOEWU(\beta)$ を計算し, $NOEWU(\beta) < mintdur$ のとき, 3 の処理を省略する.

5. 評価実験

従来手法 [1] と提案手法を実装し, 作成したオカレンスの数と実行時間を比較した. 本実験は, 報告者がリストバンド型ライフログレコーダで収集した 207 日間の運動状態データを用いて行った. $maxspan$ を 600 分, $maxgap$ を 60 分として, $mintdur$ を変化させたときに作成したオカレンスの数を図 1 に示す.

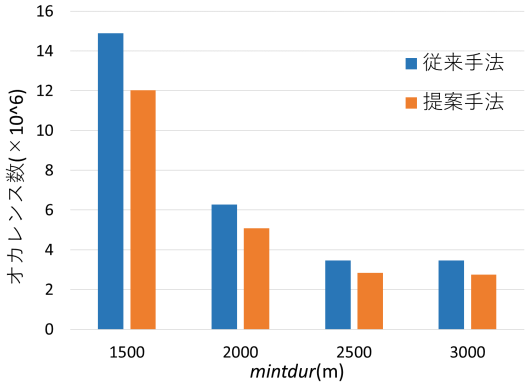

図 1: $mintdur$ を変化させた場合のオカレンス数

図 1 に示した作成したオカレンスの数に注目すると, 従来手法と比べ, 提案手法では作成したオカレンスの数が 18.2-20.4%減少していた. また, 実行時間に注目すると, 提案手法は従来手法と比べ, 実行時間が 15.8-16.2%短くなっていた. 実験結果より, $NOEWU$ による判定を従来手法に導入することにより, オカレンス作成処理を削減し, 実行時間を短縮できることを確認した.

6. まとめ

本研究では長時間エピソードマイニングを扱い, 新しい上限値を導入することによってオカレンス作成処理を削減する手法を提案した. 報告者のデータを用いた評価実験を行い, 作成されるオカレンスの数を提案手法によって低減できることを確認した.

参考文献

[1] T. Shintani, T. Ohmori and H. Fujita, Method for Comparing Long-term Daily Life using Long- duration Episodes, DARLI-AP, 2019.

[2] C. Wu, Y. Lin, Y. Philip, T.Vincent, Mining High Utility Episodes in Complex Event Sequences, ACM SIGKDD, 2013.