学術・技術論文

Partly Locality Sensitive Hashing を用いた時系列データからの高頻度パターン抽出

小川原　光　一* 田　邉　康　史* 倉　爪　　亮* 長谷川　　勉*

Detecting Frequent Patterns in Time Series Data using Partly Locality Sensitive Hashing

Koichi Ogawara*, Yasufumi Tanabe*, Ryo Kurazume* *and* Tsutomu Hasegawa*

Frequent patterns in a time series data are useful clues to learn previously unknown events in an unsupervised way. In this paper, we propose a method that detects frequent patterns in a long time series data efficiently.

The major contribution of the paper is two-fold: (1) Partly Locality Sensitive Hashing (PLSH) is proposed to find frequent patterns efficiently and (2) the problem of finding consecutive time frames that have a large number of frequent patterns is formulated as a combinatorial optimization problem which is solved via Dynamic Programming (DP) in polynomial time $O(N^{1+1/\alpha})$ thanks to PLSH where N is the total amount of data. The proposed method was evaluated by detecting frequent whole body motions in a video sequence as well as by detecting frequent everyday manipulation tasks in a motion capture data.

Key Words: Frequent Pattern Mining, Approximate Nearest Neighbor Search, Unsupervised Learning, Video Analysis

1. は　じ　め　に

ロボット技術を生活空間に積極的に導入して生活支援などに活用していくための取り組みが注目を浴びているが，そのためには周囲の人間の活動を認識するための技術が欠かせない．人間の活動の中でも高次の行動を認識する方法として，想定するタスクにおいて必要十分な行動認識器を人間が事前に設計して用いる方法 [1]〜[3] が従来提案されてきたが，生活空間における人間の活動は多様であり，タスクを限定しない場合にはこれらを網羅する行動認識器を事前に用意することは困難である．

そのため，生活空間を対象とするシステムは，新規の行動に対する認識器を逐次的に自動獲得する仕組みを持つことが望ましい．このとき，観測データ中に何度も現れる運動パターンはタスクにとって重要な意味を持つ可能性が高く，行動認識器のための学習データとして，もしくは行動文脈の学習や行動予測のための手がかりとして有用であると考えられる．そこで，本稿では上記仕組みのための基礎技術として，タスクに関する事前知識なしに観測データから頻出する未知の運動パターンを効率よく抽出する方法を提案する．

提案手法の主な特長は，(1) 局所性を保持するハッシュ関数と局所性を保持しないハッシュ関数を組み合わせた Partly Locality Sensitive Hashing (PLSH) を提案し，近似最近傍探索の枠組みによって高頻度パターンを効率よく探索する点と，(2) 非線形伸縮しかつ種類の異なる高頻度パターンの抽出問題を組合わせ最適化問題として定式化し，動的計画法を用いて全体として $O(N^{1+1/\alpha})$ の計算量で解く点の 2 点である．

以降では，まず 2 章で関連研究について述べ，3 章で提案手法の概要を述べる．次に，4 章で PLSH について説明し，5 章で時系列データから高頻度パターンを抽出する方法について説明する．最後に，6 章で提案手法の評価実験を行い，7 章で本論文をまとめる．

2. 関　連　研　究

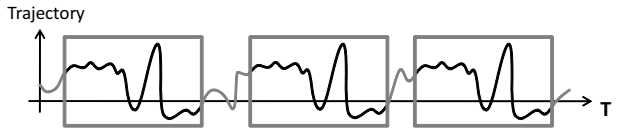

Fig. 1　Frequent patterns in a time series data

データ列の中から既知パターンの出現個所を効率よく求める方法については多くの先行研究があるが [4] [5]，本研究では Fig.1 に示すように未知の高頻度パターンの抽出を目的とする．未知

原稿受付
*九州大学
*Kyushu University

の高頻度パターンの抽出については，これまで生物情報学 [6]，データマイニング [7]〜[9]，動画解析 [10] [11]，運動解析 [12]〜[14] などの分野で精力的に取り組まれてきた．

生物情報学の分野では，塩基配列を 4 種類の塩基から構成される離散値データ列とみなし，例えば Staden らによって長さが既知の塩基配列全ての組み合わせに対して投票を行うことにより計算量 $O(N)$ で未知の高頻度パターンを抽出する手法が提案された [6]．

一方，連続値データ列に対する解析はデータマイニングや運動解析の分野で盛んに取り組まれてきた．パターン長が既知の場合は全探索の計算量が $O(N^2)$ となるため，平均計算量を下げるアルゴリズムが数多く提案されてきた．Lin らは，連続値データを離散化し，ハッシュ関数を用いた投票によって未知の高頻度パターンを効率よく抽出する手法を提案した [7]．Mueen らは，パターン同士の類似度に応じて入力データ列を並び替えることにより，連続値データのまま最も類似したパターン組を正確かつ非常に効率よく求める手法を提案した [9]．

パターン長が未知の場合には，動的計画法（ＤＰマッチング）に基づく多くの手法が提案されてきた．内田らは，論理判定型ＤＰマッチングを用いて，1 つのデータ列中に複数回出現するパターンを計算量 $O(N^2)$ で抽出する手法を提案した [13]．平均計算量を $O(N^2)$ 未満に下げるアルゴリズムとして，Yankov らはパターン長を離散的に一様伸縮し，Lin らの手法 [7] を拡張して伸縮する高頻度パターンを抽出する手法を提案した [8]．Meng らは，Locality Sensitive Hashing (LSH) [15] を用いて時刻ごとに類似データを探索しこれらを接続することによって，モーションキャプチャデータから非線形伸縮する高頻度パターンを計算量 $O(N^{1+1/\alpha})$ で抽出する手法を提案した [14]．しかし，データ数 N の増加に伴いハッシュバケット内のデータ数も増えるため，バケットの大きさが固定値の場合には実際の計算量は $O(N^2)$ に近くなる．

本稿では，主にこの [14] の問題点を解決する手法を提案する．[14] の手法では時系列データの各データ点ごとに LSH を用いた近似最近傍探索を行っている．しかし，対象が連続した時系列データである場合には，時間方向に近いデータ点における探索結果を用いて現データ点の探索結果を補うことが可能である．そこで提案手法では，各データ点においては限定された近似最近傍探索を行い，近いデータ点同士でこの限定された近似最近傍探索の探索範囲が互いに独立になるように探索空間を構築することによって，計算時間の観点から効率のよい高頻度パターン探索を実現する．

そのために，本稿では近似最近傍探索法の一種である Partly Locality Sensitive Hashing (PLSH) を提案し，LSH を利用した近似最近傍探索法と比較してより少ない計算時間で高頻度パターンが抽出できることを示す．また，提案手法では高頻度パターン抽出を組合わせ最適化問題として定式化し動的計画法を用いて解くことによって，全体の計算量を [14] と同様に $O(N^{1+1/\alpha})$ に抑えた．ただし，$\alpha(>1)$ は PLSH のパラメータによって決定される定数である．

高頻度パターンの抽出は，高頻度パターンに低ビットのコードを割り付けるデータ圧縮問題として考えることもできる．Zhao らは，最小記述長 (Minimum Description Length) 基準に基づき符合化データ列と辞書の大きさの和を最小化することによって，舞踏の運動計測データの分節化を行った [12]．しかし，低頻度パターンが支配的である通常のデータ列に対しては，このような方法は適当ではない．

3. 提案手法の概要

Fig.1 に示すように，本研究では長時間の時系列データ (d 次元) が与えられたときに，そこから高頻度で出現する類似した未知パターンを抽出することを目的とする．パターンとは時系列データの部分データ列を指し，パターン同士でパターン長や対応するデータ値の差が小さいものを類似パターンとする．

Fig.2 に，時系列データを d 次元から 2 次元に投影して表示した例を示す．もし時刻 t のデータ点 $o(t)$ がある高頻度パターン群に属している場合，このデータ点の近傍に他の類似パターンも存在することになる．つまり，あるデータ点の近傍に他の多くのパターンが存在するということは，そのデータ点が高頻度パターン群に属しているための必要条件となる．そのため，近傍パターン数の多いデータ点が連続する区間は高頻度パターンのよい候補になると考えられる．

そこで,「近傍」を d 次元空間における半径 R の超球内と定義し,「セグメント」を時系列データのうち超球に含まれる各部分データ列と定義して,「データ密度」を全セグメントの長さの合計と定義する．すると，データ点 $\boldsymbol{o}(t)$ を中心としたデータ密度は

$$D(t) = \sum_{i \in S(t)} \|\boldsymbol{o}(i) - \boldsymbol{o}(i+1)\| \quad (1)$$
$$\text{where } S(t) = \{i; \|o(i) - o(t)\| \leq R\}$$

と計算され，これを各時刻ごとに計算することによって高頻度パターンの有無を評価することを考える．このとき，もし各時刻ごとに厳密に計算を行うと，データ構造を木構造にするなどの工夫をしたとしても計算時間は N^2 に依存した値となり時間がかかる．

そこで，本稿では近似最近傍探索の枠組みを用い，Table 1 に示す方法でデータ密度を効率よく計算する．

Table 1 Algorithm to find frequent patterns

1. 時刻 $t = 1$:
   超球内のセグメントの情報を保持するリンクリストを初期化（Fig.2(a)）
2. 時刻 $t = 2$ から N:
   リンクリストを更新（Fig.2(b),(c)）
   PLSH を用いて新規セグメントを検出しリンクリストへ追加（Fig.2(d)）
   PLSH を用いて分断セグメントを検出しリンクリストを更新（Fig.2(e)）
3. 時刻 $t = 1$ から N:
   大域最適化に基づき全高頻度パターンを抽出

時刻 $t = 1$ では，N 個の全データ点を調べ，超球内のセグメントの境界（時系列データと超球面との交点），つまり各セグ