ショートノート

順序列から任意の出現頻度以上の部分列を抽出するニューラルネットワーク†

森田 賢太*¹・高瀬 治彦*¹・川中 普晴*¹・森田 直樹*²

本稿は与えられた系列データから頻出な部分列を抽出することを目的とした。特に、抽出に際して (1) オンライン学習, (2) 複数部分列の抽出, (3) さまざまな長さの部分列の抽出, (4) 頻出とするしきい値の調節の 4 つすべて可能にすることをめざした。提案手法は、スパイクニューロンを用いた 2 ブロックからなるニューラルネットワークである。LIF モデルに基づくユニットにより構成し、STDP 学習則に基づいた結合荷重の更新を行うことで、自己組織的に部分列を抽出するネットワークを構築する。この結果、1 つの STDP のパラメータを調整するだけで、同じ系列から頻出として抽出する部分列を変化させることができた。具体的には、3,000 シンボル長の系列から 3 シンボル長の部分列 (出現頻度は 0.4%, 3%, 5%) を抽出した際、3% 以上出現する部分列の抽出と 5% 以上出現する部分列の抽出の切り替えに成功した。

キーワード：頻出パターンマイニング、時系列処理、スパイクニューラルネットワーク

1. はじめに

頻出パターンマイニングはデータ集合から頻出に出現するパターンを自動的に抽出するタスクである。データ集合の対象としては、系列データやアイテム集合などがある。系列データに対する頻出パターンマイニングはただ単に頻出パターンを抽出するのではなく、順序を考慮してパターンを抽出する。例えば、系列データ内に要素 A の後に要素 B が出現するパターンが多い場合、頻出部分列として“AB”を抽出する。

文章データや音声データなどの系列データから頻出な順序パターンを抽出する場合、オンライン学習、複数の部分列の抽出、さまざまな長さの順序列パターンの抽出、頻出とみなすしきい値の調整、の 4 つのことが必要である。

系列データから頻出な順序列パターンを抽出する方法として、いくつか手法が提案されている。Sakurai らは、Apriori アルゴリズムに基づく手法を提案した [1]。この手法は、頻出アイテムを検出するアルゴリズムである Apriori アルゴリズムを拡張し、上位 k 個の非類似で頻出な順序パターンを検出できるようにしたものである。しかし、1 つ目のオンライン学習ができない。Tanaka らは、連想記憶を利用した手法を提案した [2]。この手法は、連続して頻繁に輸入される 2 つの要素を連想記憶により記憶することで、頻出な順序列パターンを検出する。この手法には抽出できる部分列の長さに制限がある。Morita らは、スパイクニューラルネットワークを利用した手法を提案した [3]。この手法は、ネットワークを構成する各スパイク

図1 従来手法のネットワークの構成

ングニューロンを用いて過去の入力情報を保持することで、頻出な順序パターンを検出する。しかし、この手法は頻出とみなすしきい値の調整のみができていない。

本稿では、系列データに対し前節で述べた 4 つができる頻出パターンマイニングの手法を提案する。前出の従来法のうち、Morita らの手法は頻出とみなすしきい値の調整のみができていない。しかしこれは、原理的にできないのではなく、調整するためにパラメータを調整することが難しいと考える。これをふまえ、我々はこの手法を拡張して頻出とするしきい値を簡単に調整できるようにすることをめざす。

2. 従来手法

本章では、今回の提案手法のもととする文献 [3] の手法について説明する。

2.1 構成

ニューラルネットワークは、図1に示すように、入力ブロックと出力ブロックの2種類のブロックから構成される。入力ブロックは、入力された要素を一定時間入力情報を保持する。入力ブロックには、入力端子と2つのユニットが鎖状に接続され

† Extraction of Frequent Sub-Sequences from Long Sequence Using Neural Network

Kenta MORITA, Haruhiko TAKASE, Hiroharu KAWANAKA, and Naoki MORITA

*1 三重大学大学院工学研究科
Graduate School of Engineering, Mie University

*2 東海大学大学院情報通信工学研究科
Graduate School of Information and Telecommunication Engineering, Tokai University

ており、各鎖は入力要素の種類に対応する。

出力ブロックは頻出の部分列を抽出する。出力ブロックは複数の層を持っており、第1層から順に、抽出する部分列の長さが2, 3, ...に対応する。第1層に属するユニットは、入力ブロックの第1層のユニットと入力ブロックの第2層の全ユニットから結合している。第2層以降に属する各ユニットは、入力ブロックの第1層のユニットと出力ブロックの1つ前の層の全ユニットから結合している。各層に属する各ユニットは、初期にはランダムな結合荷重のためランダムに発火するが、ネットワークの学習後には各ユニットが頻出部分列の1つ1つに対応し、対応する部分列が入力された時にのみ発火する。

2.2 動作

入力ブロックのユニットは、信号を受け取ると即座に発火して、接続しているユニットへ信号を伝播する。伝搬には要素の入力間隔と同じ時間の遅延が発生する。これにより、入力ブロックは入力された瞬間および1ステップ前に入力された入力要素を保持する。

一方、出力ブロックのユニットの発火動作は LIF (Leaky Integrate and Fire) モデルに基づいており、以下の数式に従い動作する。

$$I_i(t) = \sum_j W_{ij} O_j(t) \quad (1)$$

$$\begin{cases} V_i(t) \leftarrow V_r & (V_i(t) > \theta) \\ V_i(t) \leftarrow V_i(t) + (V_r - V_i(t)) \exp(-\frac{t_d}{\tau_m}) + I_i(t) & (\text{otherwise}) \end{cases} \quad (2)$$

W_{ij} はユニット j からユニット i への結合荷重の値で $O_j(t)$ は時刻 t のユニット j からの入力の有無であり、 $I_i(t)$ はそれらの総和である。 $V_i(t)$ は内部電位、 V_r は内部電位の初期値、 t_d は前回の入力からの経過時刻、 τ_m は時定数である。この内部電位がしきい値 θ を超えることにより発火し、次のユニットへ信号を伝播する。伝搬には、入力ブロックと同様要素の入力間隔と同じ時間の遅延が発生する。これにより、結合元のユニット群の1ステップ前の発火状態に呼応して発火ようになる。出力ブロックのユニットは、同じ層のユニット同士で側抑制を行う。側抑制により、各層では1つのユニットだけが発火する。結合元のユニット群が頻繁になる発火状態になったステップの次のステップで接続先のユニットが発火するように、LIF モデルを用いることで STDP 則による学習が簡単にできる。

2.3 学習方法

学習により、出力層の各ユニットは対応する部分列が入力された際にのみ発火するようにオンライン学習で自己組織化する。学習は出力ユニットに接続するすべての結合荷重を調整する。具体的な結合荷重の更新は、下の式に従い行う。

$$\begin{cases} W_{ij} \leftarrow W_{ij} + A_+ \exp(-\frac{x}{\tau_{stdp}}) & (x > 0) \\ W_{ij} \leftarrow W_{ij} - A_- \exp(\frac{x}{\tau_{stdp}}) & (x < 0) \end{cases} \quad (3)$$

A_+ と A_- は定数、 x はユニット i とユニット j の発火時刻の差

で、 τ_{stdp} は時定数である。接続元のユニットが発火した後に接続先のユニットが発火した場合は結合荷重の値が増加し、逆の順番で発火した場合には値は減少する。増減量は接続元と接続先ユニットの発火時刻の差が小さいほど大きくなる。これにより、接続元ユニット群の発火状態が頻繁に起こる状態になった後に、接続先ユニットが発火するように荷重が調整される。調整を繰り返す行うことで、頻度が高い部分列に対してのみ反応する出力ユニットを得ることができる。調整をうまく行うために、さらに以下に示す2つの工夫がされている。

1つ目は、発火しない時間に基づいて、結合荷重の値を減衰させることである。STDP 学習則はユニットが発火しなければ結合荷重の値が変化しない。なので、以下の式に従い発火頻度が低いユニットが接続元の結合荷重を減少する。

$$W_{ij} \leftarrow W_{ij} - D_- \exp(\frac{x_{exp}}{\tau_{exp}}) \quad (4)$$

D_- は定数、 x_{exp} は接続元のユニットが前回発火してからの経過時刻、 τ_{exp} は時定数である。

2つ目は、各結合荷重の値を一定範囲内に制限することである。結合荷重に最大値を設定することにより、2つの層から信号が来た時のみ出力層のユニットが発火する。学習により、ある出力層のユニットへ接続している結合荷重の1つが最大値に達した場合、その接続元であるユニットと同じ層に所属しているユニットからその出力層のユニットへの他の結合荷重の値をすべて0にする。ある出力層のユニットへ接続している2つの層からの最大値に達した結合荷重ができたとき、そのユニット1つは部分列を抽出した状態となり、特定の順序で要素がニューラルネットワークに与えられた時のみ発火ようになる。抽出した出力層のユニットへ接続している結合荷重は、そのあとニューラルネットワークに要素を与えても変化しない。この時点で残っている結合より、どの部分列を抽出したのかを判断できる。

このネットワークは、出力ブロックのユニットが部分列を抽出するたびに、ユニット及び出力ブロックの層を増やす。また、出現頻度が高く長さが短い部分列から順に、対応する出力ブロック内のユニットが決定する。このため、層数を抽出したい部分列の長さに対応した数に制限し、各層のユニットの個数を抽出したい個数に対応した数までに制限することで、抽出する部分列の長さ・数を調整する。

3. 任意の出現頻度の部分列を抽出するために

本章では、任意の出現頻度以上の部分列を抽出するために、2章で説明したニューラルネットワークに以下の変更を行う。

従来法において、学習に失敗する要因の1つはユニットの発火頻度が適正な範囲内で維持できないことであった。発火頻度が低すぎると、STDP 則で結合荷重を更新する機会が減り学習しづらくなる。発火頻度が高すぎると、STDP 則で頻繁に荷重が更新され結合荷重が適切に収束しない。各ユニットの発火頻度を決定するのは STDP 則 (式 (3)) および式 (4) である。これらの強さを決めるパラメータは A_+ 、 A_- 、 D_- であり、抽出したい出現頻度に応じてこれらすべてを適切に決めるのは困難であった。

任意の出現頻度以上の部分列を抽出するためのパラメーター設定を容易にするために、2章で説明したニューラルネットワークの学習則に以下の2点の変更を行う。一点目の変更は、式(4)を利用しないことである。式(4)は、頻度が少ない部分列に対応している結合の荷重を減衰させることで、それを忘却させるために使用されている。これを使用しないことで、この忘却が行われなくなるが、2.3節で述べた第2の工夫により、高い頻度の部分列を検出した時点でそのような結合は除かれるので問題ないとする。ただし、学習の過程で誤って頻度が低い部分列に反応した結果がいつまでも残り、誤検出につながる可能性がある。この点については、実験にて確認する。二点目の変更は、パラメーター A_+ , A_- を同じ値にすることである。これらのパラメーターは、各ユニットでの発火タイミングが正しいとき及び誤っているときの更新量を制御する。これらについては各ユニットの発火頻度には大きく関わっていないので、簡単のために同じ値にする。

ユニットの動作の変更は、ユニットの発火頻度を一定の範囲内で維持するように行う。従来手法のユニットのモデルでは、パラメーター A_+ , A_- の値により、発火の頻度が異なった。従来のユニットの内部電位は、結合荷重の値を入力として受け取り内部電位に加算し、その内部電位を入力がない時間に応じて減衰させていた。そのため、高い結合荷重による入力が続けば内部電位が上昇しやすく頻繁に発火したり、低い結合荷重による入力が続けば内部電位が上らずに発火しにくくなる。そのため、各ユニットの発火頻度を一定の範囲内で維持することが難しい。これを防ぐために、結合荷重とユニットの間に信号を受け取る受容体を用意し、その受容体ごとに値を減衰させ、ユニットの内部電位はその受容体の値の積和とする。これにより、入力刺激による内部状態への影響が、結合ごとに分割されるため、各結合での発火頻度に合わせた学習をし易くなる。具体的なユニットの内部電位の計算は、次の式に従う。

$$R_{ij}(t) \leftarrow R_{ij}(t) \exp(-\frac{t_d}{\tau_m}) + W_{ij} O_j(t) \quad (5)$$

$$\begin{cases} V_i(t) \leftarrow \sum_j R_{ij}(t) & (\text{otherwise}) \\ V_i(t) \leftarrow V_r & (V_i(t) > \theta) \end{cases} \quad (6)$$

$R_{ij}(t)$ は時刻 t の時の受容体の値である。受容体の値の積和がしきい値 θ を超えた場合発火する。発火した場合、同じ層に所属している他のユニットへ側抑制の信号を送る。側抑制の信号を受け取ったユニットは、その受け取ったタイミングに受容体に入る値は0にし、信号の伝搬はなかったこととする。

このネットワークは結合荷重が増加することで部分列を抽出することができ、結合荷重の増加量はパラメーター A により決まる。そのため、パラメーター A を大きくすれば結合荷重が増加しやすいため出現率が低い部分列を抽出することができ、逆にパラメーター A を小さくすれば結合荷重が増加しにくいため出現率が高い部分列のみを抽出できると考えている。このため、提案するニューラルネットワークはパラメーター A の調整により任意の出現率以上の部分列を抽出できると考えられる。

表1 確認用データの各部分列の出現率

部分列	出現頻度 (%)
ABC	5.0
XYZ	3.1
その他各	0.4 以下

表2 ネットワーク、学習のパラメーター

項目	値
初期の結合荷重	5.0
結合荷重の最大値	10.0
θ	-50.0
V_r	-68.0
受容体の最大値	10.0
τ_m	145.0
τ_{stdp}	10.0

表3 抽出結果

A	出力ユニット					
	第1層				第2層	
	1 番目	2 番目	3 番目	4 番目	1 番目	2 番目
0.16	AB	CA	BC	-	ABC	-
0.30	AB	YZ	XY	BC	ABC	XYZ

4. 実験

本章では、提案するニューラルネットワークが学習則のパラメーター A のみの調整により、抽出する部分列の出現率を調整できるのかを確認する。

4.1 確認方法

学習時に与える系列データは、乱数を用いて特定の部分列の出現頻度が高くなるように作成する。その系列データを日本語の音素間の平均間隔である 0.145 秒間隔で1要素ずつネットワークに与える。確認用の系列データの作成にはアルファベットの大文字と小文字の 52 種の要素を使い、データの長さは 3,000 要素とした。表1は作成した系列データの要素数3の部分列の出現頻度を表している。順序列パターン“ABC”は高頻度の部分列で、順序列パターン“XYZ”は中頻度の部分列である。学習則のパラメーター A のみを調整することで、提案するニューラルネットワークがこの確認用系列データから抽出する部分列を、高頻度の順序列のみと、中頻度以上の部分列とを切り替えることができるのかを確認する。

ネットワークと学習則に関するパラメーターは、表2のとおりにした。

4.2 抽出確認

提案するニューラルネットワークの学習則のパラメーター A を変化させて、ネットワークが抽出した部分列を調査した。なお、長さ3の部分列“ABC”が頻出であれば、その部分列である“AB”、“BC”も頻出である。提案したニューラルネットワークでは、短い部分列から順に抽出される。以下、本実験では分析を簡単にするために出力ブロックの第1層と第2層にのみ着目し、長さ3までの部分列についてのみ議論する。

表3は、ニューラルネットワークの学習則のパラメーター A をそれぞれ 0.16 と 0.30 にしたときの抽出結果を表している。パラメーター A が 0.16 のときは 5% 以上の部分列のみ、パラメーター A が 0.30 のときは 3% 以上の部分列のみが抽出されている。

この結果から、提案するニューラルネットワークはパラメーター A の調整により任意の出現率以上の部分列のみを抽出できると考えられる。

5. まとめ

本稿は、系列データから頻出な順序パターンを抽出する際にオンライン学習、複数の部分列の抽出、さまざまな長さの部分列の抽出、頻出とみなすしきい値の調整の4つができる頻出パターンマイニングをめざした。文献[3]の手法を拡張して4つができる頻出パターンマイニングの手法を提案した。ニューラルネットワークの学習則とユニットのモデルを変更することで、パラメーター A の変更のみで頻出パターンの基準となる頻出率のしきい値を調整できるようにした。簡単な実験の結果、学習則のパラメーター A を調整することで、任意の出現頻度の順序列を抽出できることを確認した。

参考文献

- [1] S. Sakurai, and N. Nishizawa: "A NEW APPROACH FOR DISCOVERING TOP-K SEQUENTIAL PATTERNS BASED ON THE VARIETY OF ITEMS," *J. of Artificial Intelligence and Soft Computing Research*, Vol.5, No.2, pp. 141-153, 2015
- [2] K. Tanaka, S. Yano, and N. Yamamoto: "A NEW APPROACH FOR DISCOVERING TOP-K SEQUENTIAL PATTERNS BASED ON THE VARIETY OF ITEMS: A neural network model for cognition of the order relation among sequential input patterns," *The Japan Society for Industrial and applied mathematics*, Vol.18, No.1, pp. 87-106, 2008
- [3] K. Morita, H. Takase, N. Morita, H. Kawanaka, and S. Tsuruoka: "A NEW APPROACH FOR DISCOVERING TOP-K SEQUENTIAL PATTERNS BASED ON THE VARIETY OF ITEMS: Stable Extraction of Frequent Sub-sequences from Sequential Symbol Input," *Procedia Computer Science*, Vol.112, pp. 1901-1908, 2017
(2018年10月2日 受付)
(2018年12月5日 採録)

[問い合わせ先]

〒514-8507 三重県津市栗真町屋町 1577

三重大学大学院工学研究科電気電子工学専攻計算機工学研究室

高瀬 治彦

TEL: 059-231-9691

E-mail: takase@elec.mie-u.ac.jp

高瀬 治彦 [正会員]

1991年名古屋大学工学部卒業。1996年同大学大学院工学研究科博士後期課程修了。1996年より三重大学工学部助手、准教授を経て現在三重大学大学院工学研究科教授。ニューラルネットワーク、教育支援システムの研究に従事。博士（工学）。日本知能情報ファジィ学会、教育システム情報学会、電子情報通信学会、IEEE 各会員。

川中 普晴 [正会員]

1999年三重大学工学部電気電子工学科卒業。2004年同大学大学院工学研究科博士後期課程修了。2004年株式会社医用工学研究所設立同社取締役。2005年NEDO養成技術者（株式会社三重TLO）三重大学社会連携フェロー。2009年同大学大学院医学系研究科修了。2006年より三重大学大学院工学研究科助手、助教授を経て現在三重大学大学院工学研究科准教授。ソフトコンピューティングとその応用、医療情報学に関する研究に従事。博士（工学）。博士（医学）。日本知能情報ファジィ学会、電気学会、日本医療情報学会、IEEE 各会員。

森田 直樹 [非会員]

1999年三重大学工学部電気電子工学科卒業。2004年同大学大学院工学研究科博士後期課程修了。2007年より東海大学情報教育センター講師。現在、東海大学情報通信学部准教授。画像認識、教育支援システムの研究に従事。博士（工学）。情報処理学会、コンピュータ利用教育学会会員。

著者紹介

森田 賢太 [学生会員]

2015年東海大学情報通信学部卒業。2017年同大学大学院情報通信学研究科修士課程修了。現在、三重大学大学院工学研究科在学中。日本知能情報ファジィ学会、コンピュータ利用教育学会、情報処理学会、IEEE 各学生会員。

Extraction of Frequent Sub-Sequences from Long Sequence Using Neural Network

by

Kenta MORITA, Haruhiko TAKASE, Hiroharu KAWANAKA, and Naoki MORITA**Abstract:**

In this paper, we aim to extract frequent sub-sequences from a given long sequence. Especially, the method satisfies following four requirements: (1) online learning, (2) extracting plural sub-sequences, (3) extracting various length sub-sequences, and (4) controlling threshold related to frequency. The proposed method uses a 2-blocks neural network. The network consists of spiking neurons based on leaky integrate and fire (LIF) model and is trained by the method based on spike timing dependency plasticity (STDP). As a result, the network extracted sub-sequences whose frequency is more than a certain threshold that is determined by only one parameter. Concretely, the network extracted three-symbols-length subsequences from 3,000 length sequence. In this case, sub-sequences appeared frequency of 0.4%, 3%, or 5%, and the network extracted 3%-and-more sub-sequences, or only 5% sub-sequences by controlling only one parameter.

Keywords: frequent pattern mining, time series information, spiking neural networkContact Address: **Haruhiko TAKASE***Computer Engineering Lab., Graduate School of Engineering, Mie University**1577 Kurimamachiya-cho, Tsu city, Mie 514-8507, Japan*

TEL: 059-231-9691

E-mail: takase@elec.mie-u.ac.jp