特徴選択に基づくA型H1N1亜型インフルエンザウイルス塩基配列の時間性および地域性の解析

On Temporal and Regional Analysis for Nucleotide Sequences of Influenza A (H1N1) Viruses Based on Feature Selection

嶋村 翔[1*]　平田耕一[1,2]
Sho Shimamura[1]　Kouichi Hirata[1,2]

[1] 九州工業大学大学院情報工学府
[1] Graduate School of Computer Science and Systems Engineering
[2] 九州工業大学情報工学研究院
[2] Department of Artificial Intelligence

Abstract: In this paper, we report temporal and regional analysis of influenza A (H1N1) viruses by using feature selection. Here, we adopt consistency-based feature selection algorithm CWC and apply it to nucleotide sequences of influenza A (H1N1) viruses.

1 はじめに

流行するインフルエンザ予測のために, インフルエンザウイルスの解析を行うことは重要な社会的課題のひとつであり, そのために, インフルエンザウイルスの塩基配列をバイオインフォマティクスや医療情報学の観点から解析することは有効である.

Makino ら [1, 2] は系統樹に基づいて塩基配列内にあるそれぞれの塩基のサイトに対し, **剪定距離** (*trim distance*) を導入した. また, Shimada ら [3] は 2009 年に起きたパンデミックの解析を行う際に, 塩基配列をクラスタリングするために剪定距離を利用している. その一方で Hamada ら [4, 5] は, **合致部分木マッピングカーネル** (*agreement subtree mapping kernel*) を含むいくつかのカーネルを利用し, パンデミック前後の塩基配列の分類, 地域性の分類, パッケージングシグナル位置の分類を行っている. これらのパンデミック前後の塩基配列の分類や地域性の分類は高い精度で行えているが, パッケージングシグナル位置の分類には成功していない. この場合, 系統樹を介して塩基配列を扱うよりも直接塩基配列を扱う方がより効果的である [4, 5].

本論文ではインフルエンザウイルス塩基配列に特徴選択を利用して解析する. ここで特徴選択とは, 機械学習における効率的な分類のために, 入力として与えられた特徴から不要な特徴を削減する手法である. 特徴選択アルゴリズムは, 大きく分けて**ランクに基づく手法** (*ranking-based approach*) と**一貫性に基づく手法** (*consistency-based approach*) がある. ランクに基づく手法は特徴ごとにクラスラベルに対する関連性を求め, より高い関連性を持つ特徴集合を選択する. 一方, 一貫性に基づく手法は特徴集合から部分集合を取り出し, 部分集合に対するクラスラベルへの関連性を求め, より高い関連性を持つ部分集合を選択する. ランクに基づく手法は特徴ごとに関連性を求めるため, 処理が高速であることが利点である. しかし, 特徴ごとに関連性を求めているため, 複数の特徴を利用してクラスラベルを決定しているような場合を考慮しないという欠点がある. 本論文では塩基配列に相関があることを想定し, 一貫性に基づく手法を採用する.

本論文で利用する特徴選択の手法は, 一貫性に基づく特徴選択である Shin ら [6, 7] の提案した **CWC**(*Combination of Weakest Components*) を採用する.

CWC の入力とするために, インフルエンザウイルスの塩基配列を特徴ベクトルとし, 塩基配列がもつ時間性と地域性に基づいてクラスを割り当て, クラスラベルとする. 時間性は 2009/4 から 2010/4 に 2010/10 を加えた一か月ごとの 14 個のクラスを割り当てる. 一方, 地域性では, 塩基配列の収集国を 6 つの地域に大別することでクラスを割り当てる. また, CWC によって選択された特徴を排除し, 再度 CWC を適用した結果についても示す.

本論文の構成は以下の通りである. まず, 特徴選択

*連絡先：九州工業大学大学院情報工学府
〒 820-8502 福岡県飯塚市川津 680-4
E-mail: shimamura@dumbo.ai.kyutech.ac.jp

と CWC について 2 章で説明し, CWC に対して与えるデータと実験内容について 3 章で説明する. CWC から得られた結果を 4 章で示し, 5 章で結果から得られた知見をまとめる.

2 特徴選択

特徴選択とは, ある特徴集合からクラスに対する関連性を持つ特徴を選択し, それ以外の特徴を除外するための手法である. N 次元整数ベクトルとしてデータが複数与えられたとき, そのベクトルを**クラスラベル** (*class label*) との関連性に従って分類することを考える. N 次元整数ベクトルの各次元を**特徴** (*feature*) といい, f_i で i 番目の特徴を示す. また, それらの特徴からなる集合 $\{f_1, ..., f_N\}$ を特徴集合といい, F で表す. 整数ベクトルとクラスラベルを合わせたものを**事例** (*instance*) といい, v で表す. 事例の持つ N 次元整数ベクトルを特徴値ベクトルといい, v_F で表し, 特徴集合の部分集合 $X \subseteq F$ に対しても同様に v_X と表す. また, 事例 v の持つクラスラベルを v_c で表す. 事例の集合を**データセット** (*dataset*) といい, S で表す.

データセット S に対して $|S|$ で S の総事例数を表す. このとき, 特徴値ベクトル v_F と同じ特徴値ベクトルをもつ事例の割合を特徴値ベクトル v_F の発生確率 $\frac{|\{u \in S | u_f = v_F\}|}{|S|}$ として $P(v_F)$ と表す. また, データセット S に対して事例 v が持つ特徴値ベクトル v_F とクラスラベル v_c が同一の事例の割合を事例 v の発生確率 $P(v_F, v_c) = \frac{|\{u \in S | u_f = v_F \wedge u_c = v_c\}|}{|S|}$ として $P(v_F, v_c)$ と表す.

本論文では一貫性に基づいた特徴選択を利用する. 特徴集合 $X \subseteq F$ が**一貫性** (*consistency*) を持つとは, X における任意の事象に対して, 特徴値ベクトルが同一ならば, それらの事例のクラスラベルが一意に決定されることをいう. これは, 以下のように定義することができる.

$$\forall u, v \in S (u_X = v_X \Rightarrow u_c = v_c)$$

しかし, データセットの特徴集合が常に一貫性を持つわけではないため, 一貫性を持つ状態に近いかの度合いとして**一貫性指標** (*consistency measure*) が用いられる. 本論文では**条件付きエントロピー** (*conditional entropy*) と**二値一貫性** (*binary consistency*) を用いる. 条件付きエントロピー $\mu_{\mathtt{ce}}$ は以下のように定義される.

$$\mu_{\mathtt{ce}}(S) = \sum_{v \in S} -P(v_F, v_c) \log \frac{P(v_F, v_c)}{P(v_F)}$$

ここで $\frac{P(f,c)}{P(f)}$ は, 特徴値ベクトルが f という値を持つとき, クラスラベルが c となる条件付き確率である.

一方, 二値一貫性 $\mu_{\mathtt{bin}}$ は以下のように定義される.

$$\mu_{\mathtt{bin}}(S) = \begin{cases} 1 & F \text{ が一貫性を持つ,} \\ 0 & \text{それ以外.} \end{cases}$$

CWC は**貪欲後方消去アルゴリズム** (*greedy backward elimination algorithm*) であり, ノイズ除去, 並び替え, 一貫性指標に基づく特徴の除外の 3 つの手順によって構成される.

3 実験手順

本論文では, **NCBI**(*National Center for Biotechnology Infomation*) が提供しているいる A 型 H1N1 亜型インフルエンザウイルス塩基配列, 2285 株を対象とし, 1 株を 1 事例とする. A 型インフルエンザウイルスは, PB2, PB1, PA, HA, NP, NA, MP, NS という **RNA 分節** (*RNA segment*) からなる. それぞれの分節の長さは 2341, 2341, 2233, 1778, 1565, 1413, 1027, 890 となっており, これらのサイトが特徴となる. ここで, 同じ塩基を持つサイトは特徴集合から除外する. 除外した結果, それぞれの特徴数は 1025, 1001, 1110, 909, 589, 682, 372, 464 となる.

時間性は収集年と月を合わせてクラスラベルとする. また, NCBI から提供されているデータには収集国で登録されているため, それらをアフリカ, アジア, ヨーロッパ, 北アメリカ, オセアニア, 南アメリカに置き換えて地域性のクラスラベルとする. 結果, 時間性は 14 クラス, 地域性は 6 クラスとなる.

表 1 はインフルエンザウイルス 17912 株に対して, 各時間性の株数とその割合 % である.

表 1: 時間性に対する菌株数と割合.

date	株数	%	date	株数	%
2009-4	1244	6.95	2009-11	2448	13.67
2009-5	2804	15.65	2009-12	2024	11.30
2009-6	2824	15.77	2010-1	544	3.04
2009-7	1568	8.75	2010-2	240	1.34
2009-8	1024	5.72	2010-3	72	0.40
2009-9	1432	7.99	2010-4	8	0.04
2009-10	1672	9.33	2010-10	8	0.04

表 2 はインフルエンザウイルス 18280 株に対して, 各地域性株数とその割合である. なお, 時間性および地域性はどちらも同一データを利用してデータセットを構成している. ただし, 地域性の情報はあるが時間性の情報がないデータが存在するため株数に差がある.

表 2: 地域性に対する菌株数と割合.

地域	株数	割合 (%)
Africa	48	0.26
Asia	1560	8.54
Europe	2792	15.27
North America	12672	69.32
Oceania	704	3.85
South America	504	2.76

4 実験

表 3 では時間性と地域性に対して選択された特徴がクラスラベルに対してどの程度の条件付きエントロピーを持つかを示している.

表 3: 選択された特徴の条件付きエントロピー.

seg	時間性	地域性
PB2	0.3117	0.057
PB1	0.3649	0.099
PA	0.3404	0.094
HA	0.3177	0.082
NP	0.6441	0.183
NA	0.5406	0.161
MP	1.1360	0.383
NS	0.8547	0.245

インフルエンザウイルス塩基配列に対し, 時間性と地域性それぞれの特徴選択を行った結果の表 3 より各分節ごとに差はあるがすべての分節において時間性に対し, 地域性が上回っている.

横軸で示す塩基配列のサイトに対して, 分節長に対する周囲 2% 範囲内に存在する選択された特徴の割合を示している. 例えば, 図 1 の PB2 では長さは 2341 であるため, 横軸で示すサイトの前後 47 のサイトに選択特徴がいくつ選ばれているかを示し, サイト 1530 ではその周囲に除外されていないサイトが 40 あり, 選択された特徴は 4 個存在するため, 0.1 となる.

図 1 および図 2 では各分節ごとに特徴が選択されているサイトに偏りがでている. 地域性に対する MP では 0 から 0.7 の範囲で増減しており, サイトごとの関連性の違いが顕著に出ている. PB2, PB1 ではあまり差が出ず, 幅としては 0.3 程度となっており, これは, クラスラベルに関連性の高いサイトが全体に散っていることを示している.

表 4 は CWC を用いて時間性および地域性に対して特徴選択を行い, 選択された特徴を除外した特徴集合を入力として繰り返し特徴選択を行った結果である. 縦

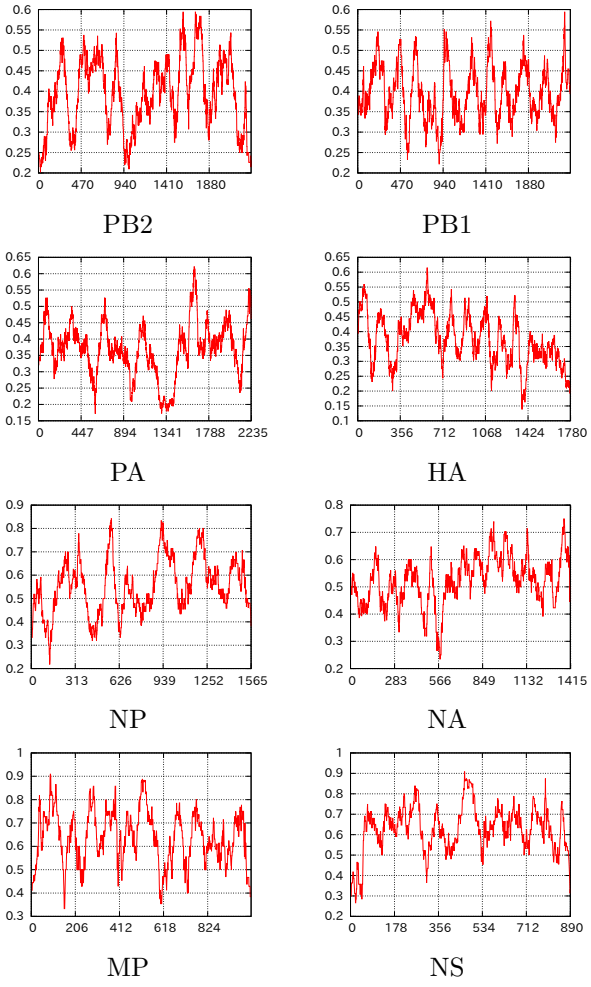

図 1: 塩基配列の時期性解析

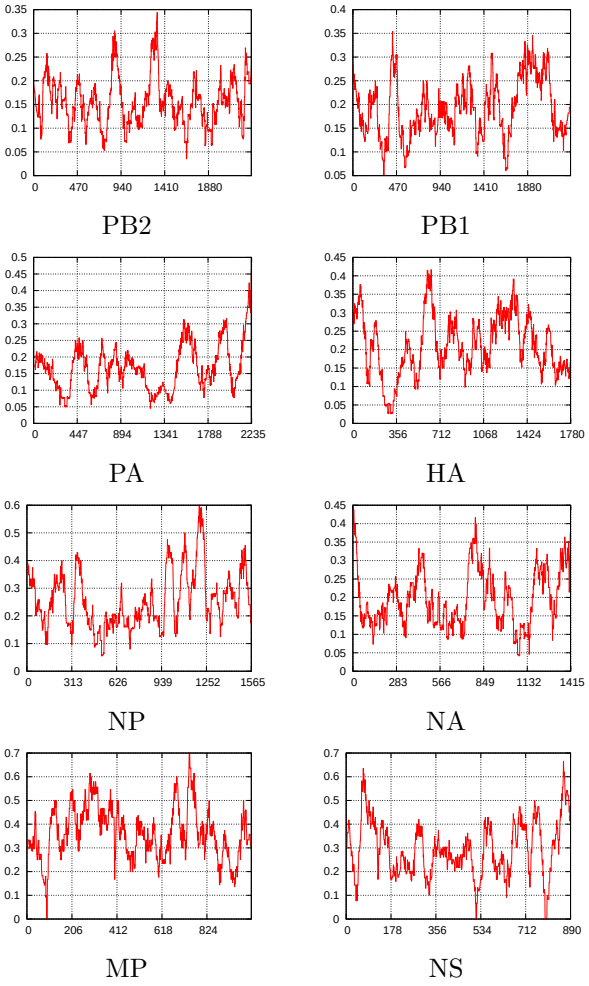

図 2: 塩基配列の地域性解析

軸は各分節を表し, 横軸は繰り返した回数を表す.

表 4: 繰り返し特徴選択を行った場合の条件付きエントロピー.

seg	01	02	03	04	05
PB2	0.0575	0.2295	0.4818	0.7134	0.8269
PB1	0.0992	0.3115	0.5528	0.6848	0.7564
PA	0.094	0.3242	0.5535	0.6869	0.7358
HA	0.0822	0.2737	0.5238	0.6877	0.7683
NP	0.1837	0.5193	0.6651	0.7652	0.7980
NA	0.1616	0.4565	0.6808	0.8369	0.8835
MP	0.3838	0.7715	0.8400	0.8614	0.8764
NS	0.2453	0.6104	0.7696	0.8583	0.8745

seg	06	07	08	09	10
PB2	0.8609	0.8674	0.8721	0.8749	0.8749
PB1	0.8567	0.8646	0.8696	0.8822	0.8822
PA	0.7924	0.7943	0.7960	0.8005	0.8009
HA	0.8204	0.8327	0.8352	0.8367	0.8367
NP	0.9006	0.9148	0.9148	0.9154	0.9154
NA	0.8873	0.8875	0.8889	0.8986	0.8986
MP	0.9118	0.9128	0.9128	0.9128	0.9128
NS	0.8942	0.915	0.9158	0.9158	0.9158

図 3 は, 表 4 の結果をグラフにしたものである. 6 回目以降から条件付きエントロピーの値がほぼ同一であったため, 7 回目までの値を記載する.

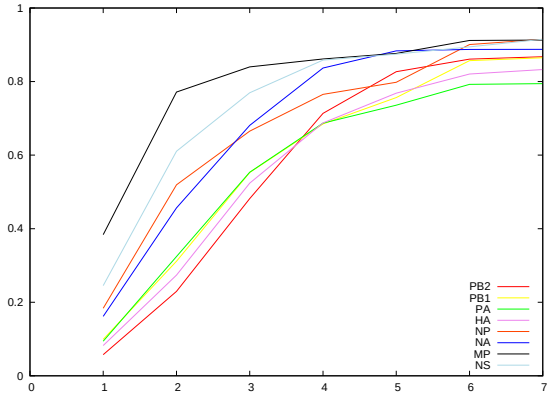

図 3: 繰り返し特徴選択を行った場合の条件付きエントロピー

繰り返し特徴選択を行った結果の図 3 から PB2, PB1, PA, HA が評価が落ちにくいことがわかる. これら 4 つの分節は地域性に対して関連性の高い塩基を多く含んでいることが考えられる.

5 まとめ

本実験の結果より, 塩基配列に対して直接機械学習で分類するのに対し, CWC を用いて特徴選択した特徴集合を用いた機械学習での地域性の分類は, 有効な手段のひとつであると考えられる. 時間性と地域性の特徴選択により選択された特徴について, 70% 程度の重複が見られた. この重複したサイトについては時間性と地域性以外のクラスラベルにも高い関連性を持つ可能性が考えられ, 他クラスへの検証は今後の課題である. また, 特徴選択により地域性, 時間性に対して高い関連性を持つサイトがあることが確認できた. このようなサイトの医学的な視点からの解析は今後の重要な課題である.

参考文献

[1] S. Makino, T. Shimada, K. Hirata, K. Yonezawa, K. Ito: *A trim distance between positions as packaging signals in H3N2 influenza viruses*. Proc. SCIS-ISIS 2012, 1702–1707, 2012.

[2] S. Makino, T. Shimada, K. Hirata, K. Yonezawa, K. Ito: *A trim distance between positions in nucleotide sequences*. Proc. DS 2012, LNAI **7569**, 1702–1707, 2012.

[3] T. Shimada, I. Hamada, K. Hirata, T. Kuboyama, K. Yonezawa, K. Ito: *Clustering of positions in nucleotide sequences by trim distance*. Proc. IIAI AAI 2013, 129–134, 2013.

[4] I. Hamada, T. Shimada, D. Nakata, K. Hirata, T. Kuboyama: *Agreement subtree mapping kernel for phylogenetic trees*, New Frontiers in Artificial Intelligence, LNAI **8417**, 321–336, 2014.

[5] I. Hamada, T. Shimada, D. Nakata, K. Hirata, T. Kuboyama: *Classifying nucleotide sequences and their positions of influenza A viruses through several kernels*,

[6] K. Shin, D. Fernalndes, S. Miyazaki: *Consistency measures for feature selection: A formal definition, relative sensitivity comparison, and a fast algorithm*, Proc. IJCAI 2011, 1491–1497, 2011.

[7] K. Shin, T. Kuboyama, T. Hashimoto, D. Shepard: *Super-CWC adn super-LCC: Super fast feature selection algorithms*, Proc. IEEE Big Data, 61–67, 2015.