反復特徴選択を用いた塩基配列の解析

On Iterative Consistency-Based Feature Selection for Nucleotide Sequences

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 propose an approach that apply the feature selection iteratively to nucleotide sequences of influenza A (H1N1) viruses. Here, we adopt an algorithm CWC (Combination of Weakest Components). The iterative application of feature selection excludes consistent sequences by the selected features until no features are consistent with the remained sequences. We extract description function from the excluding instance.

1　はじめに

本論文では, 特徴選択を利用したインフルエンザウイルスの塩基配列を解析する．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. 特徴選択の手法は様々なものがあるが, 本論文では一貫性に基づいた特徴選択を利用し, Shin ら [5, 6] の **CWC**(*Combination of Weakest Components*) を採用する.

これまでの研究で, 一貫性に基づいた特徴選択をインフルエンザウイルスの塩基配列へ適用し, 塩基の特徴について時間性と地域性の観点から解析した. [7] これにより繰り返し特徴選択を適用することでより詳しく塩基の説明ができる可能性が示唆された．そこで本論文では大きく分けて 3 つの手順で解析を行う．まず, 塩基配列にクラスラベルを付与した**データセット** (*dataset*) に対して特徴選択を行い, クラスラベルへの関連性が高い特徴を選択する．ここでクラスラベルとの関連性として一貫性指標を用いる. 次に, 特徴選択で選択された特徴を用いて一貫性を説明できる塩基配列を除外し, この時除外した塩基配列から規則性を取得する．最後に, 選択した特徴を塩基配列から除外して新しいデータセットとする．このデータセットを再度特徴選択にかけて同様の処理を繰り返すことで, より厳密なデータセットの解析を行う.

*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

CWC は入力として与えられたデータセットに対し, クラスラベルに対する一貫性を判断し, 一貫性を保てる最小の特徴を選択するアルゴリズムである. 本論文では CWC の入力とするために, インフルエンザウイルス情報から塩基配列を特徴ベクトルとし, インフルエンザウイルス情報に含まれる収集国をそれぞれ Africa, Asia, CentralAmerica, Europe, NorthAmerica, Oceania の 6 地域に分類してクラスラベルとした.

本論文の構成は以下の通りである, まず, 特徴選択と CWC について 2 章で説明し, 提案手法である特徴選択の繰り返し適応について 3 章で説明する．3 章で説明したアルゴリズムについて実験した結果を 5 章で説明し, その実験の手順を 4 章で紹介する. 6 章で得られた知見をまとめる.

2　特徴選択

特徴選択の入力として用いるデータセットとは**インスタンス** (*instance*) の集合であり, インスタンスは**特徴値ベクトル** (*feature vector*), **クラスラベル** (*class label*) で構成される．ここで, 特徴値ベクトルとは n 次元ベクトルの値を指し, $\boldsymbol{v_F} = \{f_0, \ldots, f_n\}$ で表す．この時, 特徴値ベクトルの各次元である $\boldsymbol{F} = \{0, \ldots, n\}$ をそれぞれ**特徴** (*feature*) という. 特徴値ベクトルに対してクラスタリングを行い, 分割されたクラスタそれぞれに文字情報を付与したものをクラスラベルといい, c で表す. 特徴値ベクトル $\boldsymbol{v_F}$ とクラスラベル c で構成される

インスタンスを $v(\boldsymbol{F}, c)$ と表し, インスタンスの集合をデータセットとして S で表す.

本論文では一貫性に基づいた特徴選択を利用する. データセットが**一貫性** (*consistency*) を持つとは, データセット S が持つ特徴の部分集合 f_u, f_v に対し, 任意の 2 つのインスタンス $u(f_u, c_u)$, $v(f_v, c_v)$ が以下の式を満たすと定義することができる.

$\forall u, v \in S(f_u = f_v \Rightarrow c_u = c_v)$.

この式はデータセットに対して一貫性を持つか否かを判断しており, この指標を **2 値一貫性指標** (*binary consistency measure*) と呼ぶ.

本論文で利用する CWC は**貪欲後方消去アルゴリズム** (*greedy backward elimination algorithm*) であり, 大きく分けて 3 つの手順によって構成される.

3 つの手順とはノイズ除去, ソート, 一貫性に基づいた除外である. まず, 入力されたデータセットから 2 値一貫性指標を阻害するインスタンスをノイズとして除外する. この時, 除外するインスタンスは最小となるようにする. 次に, 一貫性指標で特徴をソートする. これは次の処理でより一貫性を持ちやすい特徴を残すためである. 最後に, 各特徴をソートした順に 1 つずつ選択する. 選択している特徴以外で 2 値一貫性を持つならば選択した特徴を除外する. これらの処理終了時に残っている特徴集合が CWC での特徴選択結果である.

3 特徴選択による反復解析

提案手法について, Algorithm1 で説明する.

Algorithm 1 Algorithm

Require: *dataset* : S

$X \leftarrow S$
while $i \neq 0$ **do**
  $V \leftarrow \emptyset$
  $\boldsymbol{f_s} \leftarrow CWC(X)$
  while $v \in X$ **do**
    if $\mu(\boldsymbol{V} \cup v(\boldsymbol{f_s}, c)) = 0$ **then**
      $\boldsymbol{V} \leftarrow \boldsymbol{V} \cup v(\boldsymbol{f_s}, c)$
    end if
  end while
  $\boldsymbol{X} \leftarrow \boldsymbol{X} \setminus \boldsymbol{V}$
  while $v \in X$ **do**
    $v \leftarrow v(\boldsymbol{f} \setminus \boldsymbol{f_s}, c)$
  end while
  $i \leftarrow |\boldsymbol{V}|$
  $\boldsymbol{R} \leftarrow \boldsymbol{R} \cup \boldsymbol{V}$
end while

Algorithm1 は特徴選択を繰り返し利用することで, 厳密な条件を求める.

STEP4 でデータセット X を CWC を用いて特徴選択することにより, n 個の特徴集合 $\boldsymbol{f_s} = \{f_0, \ldots, f_n\}$ を取得する.

STEP5 からは, データセット X のすべてのインスタンス v に対し, 特徴選択で選択された特徴集合を用いて $v(\boldsymbol{f_s}, c)$ とし, インスタンス集合 $\boldsymbol{V}$ に加えて一貫性を確認する. $\boldsymbol{V}$ が一貫性を持つならばインスタンス集合 $\boldsymbol{V}$ に $v(\boldsymbol{f_s}, c)$ を加えて更新する.

STEP10 はデータセット X からインスタンス集合 $\boldsymbol{V}$ に含まれるインスタンスを除外して更新する.

STEP11 でデータセット X に含まれる各インスタンスの特徴値ベクトルから特徴 $\boldsymbol{f_s}$ を持つ特徴値ベクトルを除外する. また, $\boldsymbol{R}$ はインスタンス集合 $\boldsymbol{V}$ から構成条件 (現在の構成条件数 $+1, f_s, v_{f_s}, c$) を作成し, 加えて保持する.

除外したデータセット X に対し, 特徴選択により選択された特徴の数 $|\boldsymbol{f_s}|$ が 0 になるまで上記処理を繰り返す.

この結果得られた $\boldsymbol{R}$ がデータセットに含まれるカテゴリを説明する条件であり, 処理後の X が一貫性を持たないインスタンス集合となる.

得られた $\boldsymbol{R}$ から, クラスラベルを推定する関数 $g(v_f, v_c)$ を定める. この関数 $g(f, c)$ は以下で定義される. S が持つインスタンス $v(f_F, c)$ において $True$ となる関数である.

$v(f, c) \in S \setminus X$ であるとき

$$g(f, x) = \begin{cases} True & (x = c(f)) \\ False & (otherwise) \end{cases}$$

4 実験手順

NCBI(National Center for Biotechnology Information) が提供している A 型 (H1N1) インフルエンザウイルス情報を利用する. 取得したデータには塩基配列情報に加え, 分離日や検出地域などのヘッダ情報が含まれている.

まず取得したインフルエンザウイルス情報を各セグメントごとに分割する. 今回対象としたのは A 型インフルエンザウイルスであるため, それぞれ pb2, pb1, pa, ha, np, na, mp, ns の 8 セグメントに分割した. 次に, 各セグメントの塩基配列情報のアライメント化を行う. そして, ヘッダ情報からクラスラベルを作成しアライメント化した塩基配列と合わせることでデータセットを作成する. 本論文ではヘッダ情報から検出された国を取り出し, Africa,Asia,CentralAmerica,Europe,North America,Oceania の 6 地域に割り当てたものをクラスラベルとした.

セグメント HA に含まれるクラスラベル数上位 3 地域の Asia,Europe,NorthAmerica からデータセットを

作成した. Asia のクラスラベルが含まれるインスタンスがおよそ 4000,Eurpoe がおよそ 2200,NorthAmerica がおよそ 6000 である. 上記インスタンスから Asia インスタンス,Europe インスタンス,NorthAmerica インスタンスがそれぞれ 1-2000,1-2000,1-2000 となる地域性データセット HA1, 1001-3000,101-2100,2001-4000 となる地域性データセット HA2, 2001-4000,201-2200,4001-6000 となる地域性データセット HA3 を作成した.

作成した各データセットに対し, 提案手法で解析を行う.

実験結果に対する評価として, 各データセットに対する説明できるインスタンスの割合の算出, データセット HA1 から得られた説明関数 $g(v_f, v_c)$ を用いた他データセットへの適用を行う.

5 実験

表 1 で各データセットに対し説明できたインスタンスの割合を示す.

表 1: データセットに対する説明できるインスタンス割合.

DS	ReProc	Start(%)	End(%)
HA1	28	5284(88.07)	5412(90.20)
HA2	25	5232(87.53)	5357(89.28)
HA3	26	5272(87.87)	5418(90.30)

ここで, "DS" とはデータセット名, "ReProc" は提案手法による繰り返し特徴選択の処理終了時までの繰り返し回数, "Start" は特徴選択を 1 度行った場合説明できるインスタンス数とデータセットに対する割合, "End" は処理終了時までに取得した条件で説明できるインスタンス数とデータセットに対する割合をそれぞれ表す.

次に, データセット HA1 で作成したルールを各データセットに適用した結果を表 2 で示す.

"DS-C" は Dataset 名とそれぞれ特定のクラスラベルを持つインスタンスに注目したデータセットとなっている. HA1-A はデータセット HA1 の Asia インスタンスを持つデータセットを指す. "Ins" は"DS-C" で指定したデータセットのインスタンス数を表す. "NCI" は提案手法である繰り返し特徴選択を行う手法でも一貫性を持たないとされたインスタンスの数を表す. Num1, Num2 はそれぞれ説明関数によりクラスラベルが得られたインスタンスの数 (インスタンス全体に対する割合) と得たクラスラベルが正しいインスタンスの数 (与えたクラスラベルが正しい割合) を表す.

表 2: 他データセットへの説明関数の適用.

DS-C	Ins	NCI	Num1(%)	Num2(%)
HA1-A	2000	108	1932(96.60%)	52(97.31%)
HA1-E	2000	396	1657(82.85%)	266(83.95%)
HA1-N	2000	84	1803(90.15%)	161(91.07%)
HA1	6000	588	5392(89.87%)	479(90.78%)
HA2-A	2000	104	1384(69.20%)	254(81.65%)
HA2-E	2000	388	1896(94.80%)	287(84.86%)
HA2-N	2000	151	1136(56.80%)	472(58.45%)
HA2	6000	643	4416(73.60%)	1013(74.99%)
HA3-A	2000	72	763(38.15%)	393(48.49%)
HA3-E	2000	352	1843(92.15%)	296(83.94%)
HA3-N	2000	158	1103(55.15%)	405(63.28%)
HA3	6000	582	3709(61.82%)	1094(65.24%)

6 まとめ

実験結果よりデータセットの説明できる範囲が, 提案手法による繰り返し特徴選択を行うことで範囲が約 25%の向上がみられた. この結果から提案手法がより厳密な一貫性に基づいたルールを求めることができることが確認できた.

また, 求めた説明関数による他データセットへの適用結果について, データセットの性質を考慮しつつ考察する. Europe はルールを作成したデータセットとテストケースの重複インスタンスが 95%を占めるため当然高い認識率である. 一方, NorthAmerica はルールを作成したデータセットとテストケースの重複インスタンスは 0 である. 未知である 2000 の入力に対し, クラスラベルを全体の 55%程度付与し, 付与されたクラスラベルが正しい確率は 60%前後である.

説明関数の認識率については限定の提案手法では一貫性で得られたデータをそのまま適応しているため, 機械学習などの技術を応用し認識率の向上に努めたい.

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] 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.

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

[5] 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.

[6] 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.

[7] S. Shimamura, K. Hirata: *On Temporal and Regional Analysis for Nucleotide Sequences of Influenza A (H1N1) Viruses on Feature Selection*, Proc. 2016 International Workshop on Smart Info-Media Systems in Asia (SISA2016), 38–42. 2016.