Vol.2018-MPS-118 No.25
Vol.2018-BIO-54 No.25
2018/6/15

対称不確実性のゆらぎと一貫性に基づく特徴選択

嶋村 翔[1,a)]　平田 耕一[2,b)]

概要：一貫性に基づいた特徴選択アルゴリズムである CWC では, すべての特徴を対称不確実性の昇順に整列することで経験的に高精度な特徴を選択することができる. 本研究では, この対称不確実性の昇順の整列にゆらぎを導入することによって選択される特徴がどのように変化するかについて考察する. そして, インフルエンザウイルス A 型 H1N1 の RNA 塩基配列を対象とした実験により, 選択される特徴の数が少なくなるゆらぎが存在すること, また, ゆらぎの導入前後で学習精度が変わらない相異なる特徴集合が選択できることを示す.

キーワード：特徴選択，一貫性に基づく特徴選択，対称不確実性

the fluctuation sort in symmetric uncertainty and consistency based feature selection

Sho Shimamura[1,a)]　Kouichi hirata[2,b)]

Abstract: The consistency-based feature selection algorithm CWC can selects accurate features in experimental, by sorting all the features in increasing order of symmetirc uncertainty. In this reaserch, by introducing the fluctuation in sorting in increasing order of symmetirc uncertainty, we investigate how features are selected. Then, by the experimental results applying to nucleotide sequences of influenza A (H1N1) viruses, we evaluate that there exists a fluctuation decreasing the number of selected features and distinct sets of selected features such that one is selected without flucuation, another is selected with flucuation and they have the same accuracy.

Keywords: feature seletion, consistency based feture selection, symmetric uncertainty

1. はじめに

特徴ごとに特徴値が定まっており, 最後にクラスラベルが付加されているような特徴データからの機械学習において, 一般に, すべての特徴が均等にクラスと関連性があるわけではない. 特に, クラスとの関連性が低い特徴は, 学習精度の低下の原因となる可能性がある. そのため, 機械学習を適用する前に, 特徴データから, 学習のノイズとなり得る冗長な特徴を事前に除外することは機械学習の精度向上に有益である.

そのような手法の一つとして**特徴選択** [6], [7] がある. 特徴選択とは, ある基準にしたがって, 特徴データの特徴集合から, クラスラベルと関連性が低い特徴を除外し, より小さい特徴部分集合を選択する手法である. 特徴選択で冗長な特徴を除去することで, データの情報量をあまり落とさずにデータ量を削減することができる.

このような特徴選択には, 大きく分けてフィルタ法, 組み込み法, ラッパー法という 3 つの主要なアプローチがある. 組み込み法やラッパー法は特定の機械学習アルゴリズムの出力を最適化することを目的としている. 一方, フィルタ法は特徴選択後に利用する機械学習とは直接関係なく, 特徴データを特定の指標値に従って最適な特徴集合を選択することを目的としている. その特性から特徴を選択するた

[1] 九州工業大学大学院情報工学府
Kyushu Institute of Technology, Kawazu 680-4, Iizuka 820–8502, JAPAN
[2] 九州工業大学情報工学研究院
[a)] shimamura@dumbo.ai.kyutech.ac.jp
[b)] hirata@dumbo.ai.kyutech.ac.jp

Vol.2018-MPS-118 No.25
Vol.2018-BIO-54 No.25
2018/6/15

びに学習と分類を行うラッパー法と比べフィルタ方は高速である.

フィルタ法は探索手法と評価基準で分類でき, 評価基準は予測精度, 情報利得などがあり, 探索戦略として完全探索, ヒューリスティックス探索, 非決定性探索がある.

評価基準として一貫性を利用した特徴選択を**一貫性に基づく特徴選択** [8], [9] という. ここで一貫性とは, 特徴集合の各特徴がクラスラベルに対して矛盾しない状態からどの程度乖離しているか示す指標である.

本研究では一貫性に基づく特徴選択であり, 探索戦略にヒューリスティックス探索を用いた **CWC**(Combination of Weakest Components)[1], [2], [3] に着目する. CWC では, まず, 特徴データが一貫性を持つように特徴データからデータを除外した後, すべての特徴を**対称不確実性** (symmetric uncertainty) の昇順で整列する. そして, 整列した特徴を順に, その特徴が除外されても特徴データが一貫性を持つならば除外し, 一貫性を持たなければ何もせず次の特徴を調べる, という処理を最後まで繰り返す特徴選択手法である.

この特徴の整列を対称不確実性の昇順にすることで, 冗長な特徴選択を避け, 特徴間で関連性の高い特徴を選択することが経験的に可能である. しかし, 対称不確実性による昇順の整列に理論的保証は与えられていない.

また, 対称不確実性は特徴単体のクラスとの関連性であり, 特徴間の相互作用を反映することができない, そのため, 特徴を対称不確実性の昇順で確認した場合, 相互作用を持つ特徴がどちらも対称不確実性が高い場合は問題ないが, 対称不確実性の低い特徴と高い特徴で相互作用がある場合には, 低い特徴が除外される可能性が高い. その結果, 相互作用が失われ, 特徴選択後の特徴数の増加や学習のノイズが発生することが考えられる.

そこで, 本研究ではアルゴリズム CWC が特徴を対称不確実性の昇順の整列にゆらぎを与える. ここでゆらぎとは, 対称不確実性昇順に並べた特徴を一部別の基準で再度整列することを指す. それにより, 対称不確実性の低い特徴の除外される優先度を下げ, 対称不確実性性の低い特徴と高い特徴の相互作用を持つ特徴部分集合を抽出する.

今回の実験では, ゆらぎを与えた整列を用いて抽出した特徴部分集合の要素数や選択された特徴, また, それを用いた機械学習の精度からゆらぎの影響を考察する.

本論文は以下のように構成される. 第 2 節では, アルゴリズム CWC で必要になる情報エントロピーと相互情報量を導入し, 一貫性に基づく特徴選択について説明する. 第 3 節では, 対称不確実性と, 対称不確実性の整列に与えるゆらぎを導入し, 本研究で使用するアルゴリズム CWC について説明する. 第 4 節では, アルゴリズム CWC で対称不確実性の昇順に整列した特徴集合に対してゆらぎを与えた実験結果と考察について述べる. 第 5 節では, まとめ, および, 今後の課題について述べる.

2. 一貫性に基づく特徴選択

本節では, 特徴選択で対象となるデータセット, アルゴリズム CWC[1], [2], [3] で必要になる情報エントロピー, 相互情報量を導入する. また, 一貫性に基づく特徴選択について説明する.

本論文では $m \times (n+1)$ の自然数の行列をデータセット D と呼ぶ. また, データセットの i 行目を $v_i = [v_{i1}, \ldots, v_{in}, v_{i(n+1)}]$ と表し, 各行をインスタンスと呼ぶ. インスタンスに含まれる $v_{i1}, \ldots, v_{in}$ を特徴値ベクトル, $v_{i(n+1)}$ の値をクラスラベルと呼び, クラスラベルは v_{ic} で表す. さらに, データセット内に出現する特徴値ベクトルの値とクラスラベルの値の集合をそれぞれ X,C で表す.

また, 各列のラベルをそれぞれ特徴とし, 1 から n までの文字列ベクトルを $F = [f_1, \ldots, f_n]$ と表し, 特徴列と呼び, データセットは特徴列を一つだけ持つ.

2.1 情報エントロピーと相互情報量

特徴 f で自然数 x が出現する経験的確率を $P(f = X)$, 特徴 f で自然数 x が出現し, かつクラスラベル v_c が y となる経験的確率を, $P(f = x, v_c = y)$ とする. 特徴 f, クラスラベル v_c において, ある自然数が出現する不確実性を示す**情報エントロピー** H を以下のように定義する.

$$H(f) = -\sum_{x \in X} P(f = x) \log P(f = x).$$

$$H(v_c) = -\sum_{y \in C} P(v_c = y) \log P(v_c = y).$$

X, C の種類が多いほど情報エントロピーは大きな値をとる.

次に, 単一の特徴 f とクラスラベル v_c の関連性を示す**相互情報量** $MI(Mutual\ Information)$ を以下のように定義する.

$$MI(f, v_c) = \sum_{x \in X} \sum_{y \in C} P(f = x, v_c = y) \log \frac{P(f = x, v_c = y)}{P(f = x)P(v_c = y)}.$$

例 2.1 表 1 のデータセットについて考える. この特徴集合 F は $\{f_1, f_2, f_3, f_4, f_5\}$ であり, インスタンス集合は $\{v_1, v_2, v_3, v_4, v_5, v_6, v_7\}$ である.
f_3 の情報エントロピーと相互情報量を求める. このとき, X は $0, 1$, C は $0, 1, 2$ である. f_3 において, 0 を取り得る確率は $\frac{2}{7}$, 1 を取り得る確率は $\frac{5}{7}$ となる. C において, 0 を取り得る確率は $\frac{5}{7}$, 1 を取り得る確率は $\frac{1}{7}$, 2 を取り得る確率は $\frac{1}{7}$ となる. したがって, 情報エントロピー $H(f_3), H(C)$ は以下の通りになる. $H(f_3) = -\left(\frac{2}{7}\log\frac{2}{7} + \frac{5}{7}\log\frac{5}{7}\right) \fallingdotseq 0.9777$. $H(C) = -\left(\frac{5}{7}\log\frac{5}{7} + \frac{1}{7}\log\frac{1}{7} + \frac{1}{7}\log\frac{1}{7}\right) \fallingdotseq 1.1488$. 同様に, 相互情報量 $MI(f_3, C)$ は以下の通りになる.

Vol.2018-MPS-118 No.25
Vol.2018-BIO-54 No.25
2018/6/15

表 1　例 2.1 のデータセット

データ	f_1	f_2	f_3	f_4	f_5	C
v_1	1	0	1	0	1	0
v_2	0	1	1	1	0	0
v_3	0	1	1	1	0	0
v_4	0	1	1	1	1	1
v_5	0	1	0	1	0	2
v_6	0	1	0	1	0	0
v_7	0	1	1	1	0	1

$$\begin{aligned} MI(f_3,C) &= \frac{1}{7}\log\frac{(\frac{1}{7})}{(\frac{2}{7})(\frac{5}{7})} + \frac{0}{7}\log\frac{(\frac{0}{7})}{(\frac{2}{7})(\frac{1}{7})} + \frac{1}{7}\log\frac{(\frac{1}{7})}{(\frac{2}{7})(\frac{1}{7})} \\ &\quad + \frac{4}{7}\log\frac{(\frac{4}{7})}{(\frac{5}{7})(\frac{5}{7})} + \frac{1}{7}\log\frac{(\frac{0}{7})}{(\frac{5}{7})(\frac{1}{7})} + \frac{0}{7}\log\frac{(\frac{0}{7})}{(\frac{5}{7})(\frac{1}{7})} \\ &\fallingdotseq 0.577. \end{aligned}$$

2.2　一貫性に基づく特徴選択

一貫性の定義は以下のように定義される.

定義 2.1　特徴集合 F に含まれる任意の特徴 f_x について, クラスラベル v_c における経験確率分布 $P(f=f_x, C=v_c)$ が 1 または 0 であるとき, 特徴集合は**一貫性がある**という.

言い換えれば, 特徴集合に含まれる特徴が同一であるときクラスラベルも同一である確率が 0 または 1 であるといえる.

一貫性に基づいた特徴選択では, 特徴集合が一貫性のある状態からどの程度乖離しているかを示す**一貫性指標** μ は以下のように定義される.

定義 2.2　(決定性) 一貫性指標 μ が特徴集合に対して以下を満たすとき, μ は決定性を持つという.
F に一貫性があるとき, そのときに限り, $\mu(F)=0$ である.

定義 2.3　(単調性) 一貫性指標 μ が特徴集合 F, G に対して以下を満たすとき, μ は単調であるという.
$F \subseteq G$ ならば, $\mu(F) \geq \mu(G)$ である.

一貫性指標を用いた特徴選択を**一貫性に基づく特徴選択**といい, 一貫性に基づく特徴選択では, 一貫性指標が一定の値以下である特徴部分集合のうち, より小さい特徴部分集合を求めることを目標としている.

本研究で用いるアルゴリズム CWC では, 一貫性指標のとして, **二値一貫性指標** Bn を利用する. 特徴集合 F の二値一貫性指標を $Bn(F)$ と表し, 以下のように定義する.

$$Bn(F) = \begin{cases} 0 & F \text{ が一貫性を持つとき} \\ 1 & \text{それ以外のとき} \end{cases}$$

3.　対称不確実性の整列にゆらぎを用いた特徴選択

本節では, 一貫性に基づく特徴選択の高速なアルゴリズムの一つである CWC[1], [2], [3] を紹介し, 対称不確実性や, そのアルゴリズムの中で与える対称不確実性の整列へのゆらぎを導入する.

3.1　対称不確実性

相互情報量をそのまま特徴選択の指標値として利用する場合, 相互情報量がより大きな値を持つ特徴に重みづけされているため, クラスとの関連性が低い場合でも, データセット全体に対して多く含まれる値が優先される. そこで, 特徴 $f \in F$ とクラスラベル C に対して, 相互情報量を正規化した f の C に対する対称不確実性 $SU(Symmetric\ Uncertainty)$ を以下の式で定義する.

$$SU(f,C) = \frac{2MI(F,C)}{H(F)+H(C)}.$$

例 3.1　表 2 のデータセットについて考える. このとき, f_2 の対称不確実性を求める. まず, 情報エントロピー $H(f_2)$, $H(C)$ を求める.

表 2　例 3.1 のデータセット

データ	f_1	f_2	f_3	f_4	f_5	C
v_1	0	1	1	0	0	0
v_2	0	1	1	1	1	0
v_3	0	1	0	1	1	0
v_4	0	1	0	1	0	1
v_5	1	0	0	1	0	2
v_6	0	0	0	1	1	0
v_7	0	1	1	1	1	0

$$H(f_2) = -\left(\frac{2}{7}\log\frac{2}{7} + \frac{5}{7}\log\frac{5}{7}\right) \fallingdotseq 0.9777.$$

$$H(C) = -\left(\frac{5}{7}\log\frac{5}{7} + \frac{1}{7}\log\frac{1}{7} + \frac{1}{7}\log\frac{1}{7}\right) \fallingdotseq 1.1488.$$

次に, 相互情報量 $MI(f_2,C)$ を求める.

$$\begin{aligned} MI(f_2,C) &= \frac{1}{7}\log\frac{(\frac{1}{7})}{(\frac{2}{7})(\frac{5}{7})} + \frac{0}{7}\log\frac{(\frac{0}{7})}{(\frac{2}{7})(\frac{1}{7})} + \frac{1}{7}\log\frac{(\frac{1}{7})}{(\frac{2}{7})(\frac{1}{7})} \\ &\quad + \frac{4}{7}\log\frac{(\frac{4}{7})}{(\frac{5}{7})(\frac{5}{7})} + \frac{1}{7}\log\frac{(\frac{0}{7})}{(\frac{5}{7})(\frac{1}{7})} + \frac{0}{7}\log\frac{(\frac{0}{7})}{(\frac{5}{7})(\frac{1}{7})} \fallingdotseq 0.5774. \end{aligned}$$

したがって, 対称不確実性 $SU(f_2,C)$ は以下の通りになる.

$$SU(f_2,C) = \frac{2MI(f_2,C)}{H(f_2)+H(C)} \fallingdotseq 0.5430.$$

同様に, $SU(f_1,C) \fallingdotseq 0.6799$, $SU(f_3,C) \fallingdotseq 0.5101$, $SU(f_4,C) \fallingdotseq 0.0873$, $SU(f_5,C) \fallingdotseq 0.4400$ となる. よって, $SU(f_4,C) < SU(f_5,C) < SU(f_3,C) < SU(f_2,C) < SU(f_1,C)$ となる. したがって特徴 f_1, f_2, f_3, f_4, f_5 を対称不確実性の昇順で整列すると f_4, f_5, f_3, f_2, f_1 となる.
表 3.1 より, 特徴の値が 1 のとき, クラスラベル 0,1,2 すべてを取る f_4 が, 特徴集合 F のうち最も低い対称不確実性の値となる. また, 特徴の値が 0 のとき, 取り得るクラスラ

Vol.2018-MPS-118 No.25
Vol.2018-BIO-54 No.25
2018/6/15

ベルの種類が 0 または 1 のみの f_1 が, 最も高い対称不確実性の値となる.

対称不確実性が大きいほど, クラスラベルとの一貫性が高い特徴となる. 対称不確実性に昇順に基づいた整列を用いることでより特徴数の少ない特徴部分集合を出力したり, 特徴選択後のデータセットを使用することで機械学習の精度を向上させることが経験的に可能となる. しかし, この整列が特徴選択に適している, という理論的保証は与えられてない.

3.2 アルゴリズム CWC

アルゴリズム 1 に, アルゴリズム CWC を示す.

input : 特徴集合 F を含むデータセット
output: 特徴部分集合 X

データセットノイズ除去をする.
F の要素を対称不確実性の昇順でソートし, 列 $\mathbf{F}$ とする.
$X \leftarrow F$;
foreach $F \in \mathbf{F}$ **do**
  if $Bn(X\backslash\{F\}) = 0$ **then**
    $X \leftarrow X\backslash\{F\}$;
  end
end
return X;

アルゴリズム 1: CWC

アルゴリズム CWC では, 入力をデータセット, 出力を一貫性のある特徴部分集合とする. まず, データセットの特徴集合 F が一貫性を持つようにノイズ除去する. データセットの特徴集合 F が一貫性を持たないとき, $Bn(F) = 1$ となる. このとき, 特徴集合 F の部分集合 $\mathcal{G}$ は定義 2 より $Bn(\mathcal{G}) \geq 1$ となるため, $Bn(\mathcal{G}) = 0$ を満たさない. このままだと二値一貫性指標を用いて特徴選択できないため, ノイズ除去する. また, ノイズ除去することで, 時間効率が上昇する.

ここで, データ a, b において, a の特徴集合と b の特徴集合が同じであるが, a と b のクラスラベルが違うデータを, 特徴集合に関して**不一致**という. 不一致な例の内, クラスラベルの数が最も多いものを持つデータ以外を**マイナー**とする. マイナーなデータを除去した後, 不一致なデータが残っている場合, そのデータを除去する.

例 3.2 表 3 のデータセットを考える.

表 3.2 では, v_1, v_2, v_5 と v_3, v_4 がそれぞれのデータ間で不一致となる. さらに, v_1, v_2, v_5 ではクラスラベル 0 が 2 つ, 1 が 1 つのため, クラスラベル 1 をもつ v_2 はマイナーとなる. ここからデータノイズ除去をする. まず, マイナーなデータである v_2 を除去する. この時点で, v_1, v_5 は不一致なデータではなくなる. 次に, 不一致なデータである v_3, v_4 を除去し, データセットが表 3.3 のようになり, データノイズ除去が完了する.

表 3 例 3.2

データ	f_1	f_2	f_3	f_4	f_5	C
v_1	0	0	1	1	1	0
v_2	0	0	1	1	1	1
v_3	0	1	0	1	0	0
v_4	0	1	0	1	0	1
v_5	0	0	1	1	1	0
v_6	1	1	1	0	0	1

表 4 データノイズ除去の結果

データ	f_1	f_2	f_3	f_4	f_5	C
v_1	0	0	1	1	1	0
v_5	0	0	1	1	1	0
v_6	1	1	1	0	0	1

次に, 特徴集合の要素を対称不確実性 SU の昇順に整列する. 整列したものを列 $\mathbf{F}$ とする.

最後に, 二値一貫性指標を用いて一貫性評価をする. 対称不確実性に従って整列した列 $\mathbf{F}$ の $i(1 \leq i \leq |\mathbf{F}|)$ 番目の特徴を F_i とし, $X = F$ とする. 対称不確実性に従った順番で $Bn(X\backslash\{F_i\})$ の値を見る. $Bn(X\backslash\{F_i\}) = 0$ となるときは $X \leftarrow X\backslash\{F_i\}$ とし, $Bn(X\backslash\{F_i\}) = 1$ となるときは X に対して何も操作せずに, i を次の値にする. すべての $i(1 \leq i \leq |\mathbf{F}|)$ の範囲で探索が完了したとき, 特徴部分集合 X を出力とする.

例 3.3 ノイズ除去済みで, 特徴は f_1 から対称不確実性の昇順に整列している表 3 のデータセットを考える.

表 5 例 3.3 のデータセット

データ	f_1	f_2	f_3	f_4	f_5	C
v_1	0	0	1	1	0	0
v_2	1	1	1	1	0	0
v_3	1	1	0	1	0	0
v_4	1	0	0	1	0	1
v_5	1	0	0	0	1	2
v_6	1	1	0	0	0	0
v_7	1	1	1	1	0	0

ここで, $X = f_1, f_2, f_3, f_4, f_5$ とする. CWC は対称不確実性の昇順にみていくので, 最初に $X\backslash\{f_1\}$ をみる. このとき, $Bn(X\backslash\{f_1\}) = 0$ より, $X \leftarrow X\backslash\{f_1\}$ とし, $X = \{f_2, f_3, f_4, f_5\}$ となる. 同様に, $Bn(F\backslash\{f_2\}) = 1$ より, X に対して操作しない. $Bn(F\backslash\{f_3\}) = 1$ より, X に対して操作しない. $Bn(F\backslash\{f_4\}) = 1$ より, $X \leftarrow X\backslash\{f_4\}$ とし, $X = \{f_2, f_3, f_5\}$ となる. $Bn(F\backslash\{f_5\}) = 1$ より, X に対して操作しない. したがって, 出力結果は $X = \{f_2, f_3, f_5\}$ となる.

Vol.2018-MPS-118 No.25
Vol.2018-BIO-54 No.25
2018/6/15

3.3 対称不確実性のゆらぎ

本研究では, 特徴集合を対称不確実性の昇順で整列した特徴の列 $\mathbf{F}$ に対して 4 つのゆらぎを以下のように定義する. なお, i は $(0 < i \leq 1)$ であるとする.

定義 3.1 (部分降順 FL_i^R) 特徴集合を対称不確実性の昇順で整列した特徴列 $\mathbf{F}$ に対して, 全特徴数 $|F|$ に対し $\mathbf{F}$ の先頭から, $i|F|$ 個ずつ組を作り, 組の中で対称不確実性の降順にする操作をゆらぎ FL_i^R という.

部分降順は, i が小さい場合, 全体として対称不確実性の昇順で並びつつも組の部分だけを逆にすることで, 局所的に選択する特徴に影響を与えることを想定する. i が 1 の場合は全体を対称不確実性の降順で整列した結果と同一になる.

定義 3.2 (部分移動 FL_i^M) 特徴集合を対称不確実性の昇順で整列した特徴列 $\mathbf{F}$ に対して, 全特徴数 $|F|$ に対し $\mathbf{F}$ の先頭から, $i|F|$ 個の特徴を $\mathbf{F}$ の列の末尾に対称不確実性の昇順のまま追加し, その後, 先頭から $i|F|$ 個の特徴を除外することをゆらぎ FL_i^M という.

部分移動は, 対称不確実性の低い特徴を一部列の後ろに移動することで, 移動した特徴を除外されにくくし, 対称不確実性の低い特徴との相互依存を残すことを想定する. i が 1 の場合は, 対称不確実性の昇順で整列した特徴列のままになる.

定義 3.3 (未選択特徴優先 FL_{CWC}^S) 特徴集合を対称不確実性の昇順で整列した特徴列 $\mathbf{F}$ に対して, アルゴリズム CWC で特徴選択を行い, 選択された特徴を特徴列 $\mathbf{F}$ の先頭に移動させ, 選択された特徴を対称不確実性の昇順, 選択されていない特徴を昇順で整列するゆらぎを $FL_{CWC}^{S_{ii}}$, それぞれ昇順, 降順で整列するゆらぎを $FL_{CWC}^{S_{id}}$, それぞれ降順, 昇順で整列するゆらぎを $FL_{CWC}^{S_{di}}$, それぞれ降順, 降順で整列するゆらぎを $FL_{CWC}^{S_{dd}}$ という.

未選択特徴優先は, 一度特徴選択を行った結果を用いて, 除外されなかった特徴を除外されやすいようにし, 別の特徴集合を取得できることを想定する.

定義 3.4 (先頭特徴優先 FL_{CWC}^M) 特徴集合を対称不確実性の昇順で整列した特徴列 $\mathbf{F}$ に対して, アルゴリズム CWC で特徴選択を行い, 選択された特徴のうち, 対称不確実性が最小の特徴までの特徴を $\mathbf{F}$ の列の末尾に対称不確実性の昇順のまま追加し, その後, 先頭から $i|F|$ 個の特徴を除外することをゆらぎ FL_i^M という.

先頭特徴優先では, 一度特徴選択を行い, 選択された特徴の中でも対称不確実性の低い特徴までの特徴を除外されにくい列の最後尾に移動させる. それにより, 選択済み特徴の対称不確実性が最も低い特徴は最も除外されやすくなり, 別の特徴集合を取得できることを想定する.

これら 4 つのゆらぎは図 1 のように表現できる. 未選択特徴優先, 先頭特徴優先では事前処理が必要であるため, 事前処理の結果を図右上に示す. 黒い部分が特徴選択を行った結果, 選択された特徴を表す. 各ゆらぎは x 軸が整列特徴, y 軸が対称不確実性を表す. 部分降順, 部分移動では $i = 0.3$ の結果である.

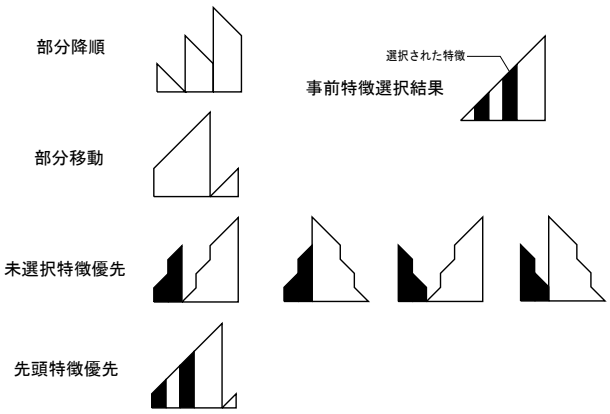

図 1 各ゆらぎ適用後の対称不確実性

4. 実験と考察

本節では, NCBI[10] から提供されている A 型 (H1N1) 型インフルエンザウイルスの 8 本の RNA 分節に対してアルゴリズム CWC での選択した特徴とゆらぎを加えたアルゴリズムで選択した特徴を比較する.

各文節の特徴数は PB2 は 2725, PB1 は 2650, PA は 3817, HA は 2358, NP は 1636, NA は 1775, MP は 1223, NS は 992 であり, それぞれのインスタンス数は 12357, 12251, 12263, 31908, 12690, 24827, 17889, 12958 となっている. また, 対称不確実性の昇順ソートで選択された特徴数はそれぞれ 545, 547, 561, 790, 502, 688, 504, 527 である. 選択した特徴の評価指標として, ゆらぎの結果を用いた機械学習 (SVM) の結果を利用している. SVM では各分節のデータセットからアルゴリズム CWC で選択された特徴のみのデータと, ゆらぎを適用した特徴選択で選択された特徴のみのデータを作成し, それぞれを 10 交差法にてトレーニングとテストを行った.

図 2 は部分降順ゆらぎの結果, y 軸で増加特徴数 $\frac{FL_i^R - FL_{CWC}}{FL_{CWC}}$, x 軸で i を 0.05 から 1 まで, 0.05 ごとに 1 回, 合計 20 パターンを示す. x 軸の値が 0 であれば, 通常の CWC で選択した特徴と同数の特徴が選択されている. 図 3 はゆらぎを適用した特徴選択結果を用いたクラス分類の精度を示す.

特徴数は i が増加するほど増加する傾向にあり, 全ての分節で $i = 1$, つまり対称不確実性の降順の際に最大となっている. 一方, クラス分類の精度はそれぞれ $i = 0.2, 0.1, 0.45, 0.1, 0.1, 0.05, 0.4, 0.55$ の時に最大になっており, i や増加特徴数と直接的な関係は見られない. 選択した特徴数とクラス分類の精度の観点からそれぞれの分節と閾値で $8 \times 20 = 160$ パターンのうちどちらも増加した結

Vol.2018-MPS-118 No.25
Vol.2018-BIO-54 No.25
2018/6/15

果が 64 パターン, 選択した特徴数は増えたがクラス分類の精度が落ちた結果が 79 パターン, 選択した特徴数が減ったがクラス分類の精度が上がった結果が 7 パターン, どちらも減少した結果が 10 パターンであった. 部分降順ゆらぎでは i が増加すると, 特定の特徴の後ろにある対称不確実性が低い特徴数が増加する. 今回の結果より部分降順ゆらぎで i が増加すると選択される特徴数が増加する傾向があるが, クラス分類の精度には大きな影響を及ぼさない. また, 部分降順ゆらぎを適用した結果選択された特徴のうち, 対称不確実性の昇順で整列した CWC の結果には含まれない新たに選択された特徴は, 最大で $i = 1$ のとき, PB1 が 26.15% となっており, 566 の特徴のうち 148 の特徴が新規特徴であった.

この結果より部分降順ゆらぎの閾値は, 特徴数に強く影響し, 選択された個々の特徴の関連性は高いが, 選択された特徴集合全体を機械学習にかけた場合のクラス分類の精度にはほぼ影響がないといえる. さらに, 対称不確実性の昇順に並べた場合と比べ新規特徴を取得したい場合は, 全体を降順で並べることで多くの新規特徴を得ることができる.

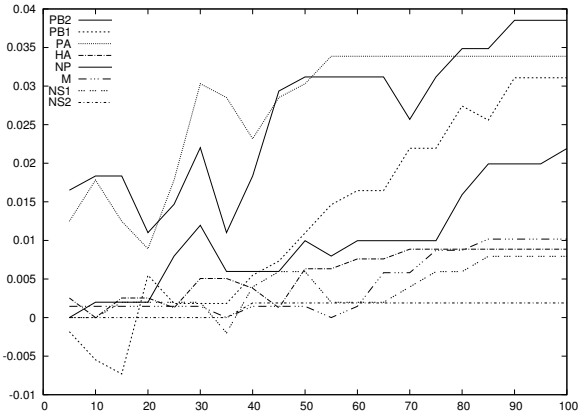

図 2 FL_i^R で選択された特徴数の変化

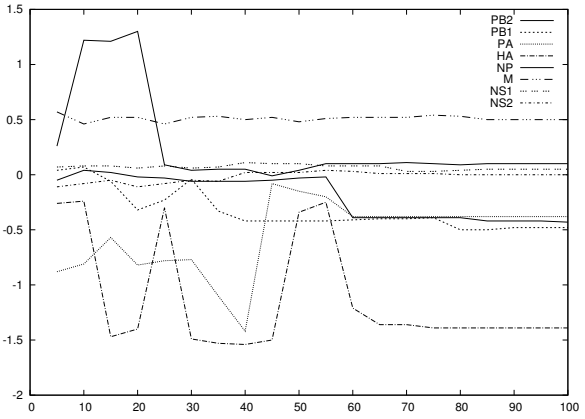

図 3 FL_i^R で選択された特徴でのクラス分類精度

図 4, 図 5 は図 2, 図 3 と同様に部分移動ゆらぎの増加特徴数と, クラス分類の精度の変化を示す. 部分移動ゆらぎでは, 全体の 20% から 30% 程度を移動させたとき最も特徴数に変化があり, 50% を超えたあたりから変化が少なくなる. クラス分類の精度は, 部分降順ゆらぎの結果と同様大きく変化していないが, 部分降順ゆらぎとは異なり, 各セグメントごとに特色が見られた. 主に, PA が 20% から 30% で認識率が低下し, HA は 50% から 70% で認識率が低下している. これはその分節における重要な特徴がその周辺にあるからではないかと考えられる. また, 部分移動ゆらぎを適用した結果選択された特徴のうち, 対称不確実性の昇順で整列した CWC の結果には含まれない新たに選択された特徴は, 最大でも 16.75% となっている.

この結果より部分移動ゆらぎは部分降順ゆらぎと同様, 得られた特徴集合全体のクラス分類精度にはあまり影響はなく, 新規特徴の取得にはあまり向いていない. しかし, 各データセットの特性を見ることに向いているといえる.

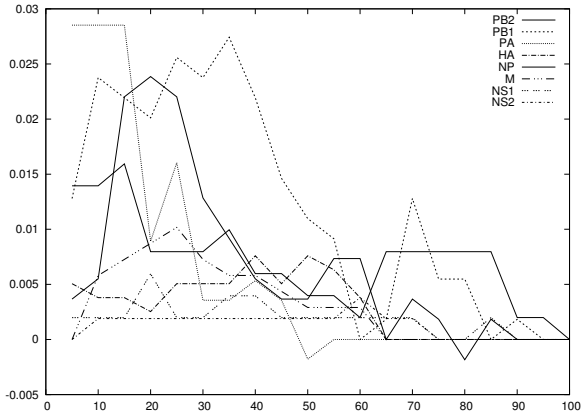

図 4 FL_i^M で選択された特徴数の変化

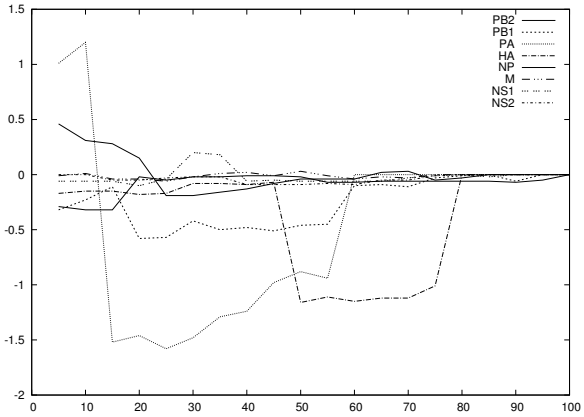

図 5 FL_i^M で選択された特徴でのクラス分類精度

図 6, 図 7 は, 特徴選択結果を用いたゆらぎである未選択特徴優先ゆらぎと先頭特徴優先ゆらぎの増加特徴数とクラス分類の精度変化を示す. x 軸は ii は $FL_{CWC}^{S_{ii}}$, id は $FL_{CWC}^{S_{id}}$, di は $FL_{CWC}^{S_{di}}$, dd は $FL_{CWC}^{S_{dd}}$, M は FL_{CWC}^{M} を

Vol.2018-MPS-118 No.25
Vol.2018-BIO-54 No.25
2018/6/15

それぞれ示す．未選択特徴優先ゆらぎでは ii, id, di, dd それぞれで選択された特徴数はほぼ同じだが, 部分降順ゆらぎや部分移動ゆらぎと比べ, 増加特徴数が 2 倍程度多い結果となった．その結果, 認識精度の向上に繋がっている．更に, 図 8 で示したように, PB2, PB1, PA については選択された特徴の 3 割程度が新規特徴となっており, 部分降順ゆらぎで得られたよりも多くの新規特徴が得られた．先頭特徴優先ゆらぎについては選択された特徴やクラス分類の精度に対してほぼ影響を与えていない．

この結果より未選択特徴優先ゆらぎは新規特徴を得るために利用できると考えらえるが, 特徴増加量やクラス分類の精度についてはデータセットに依存する部分が大きい．また, 先頭特徴優先ゆらぎの結果より, 対称不確実性の昇順に並べて選択したとき, 最小の対称不確実性を持つ選択特徴より対称不確実性が低い特徴は, 除外の優先度に関わらず, 除外される可能性が高いといえる．

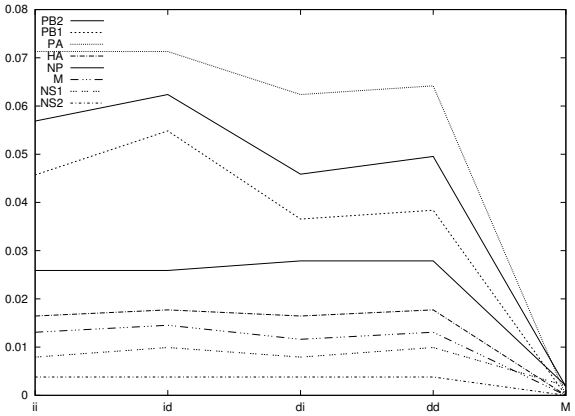

図 6 特徴選択結果を用いたゆらぎで選択された特徴数の変化

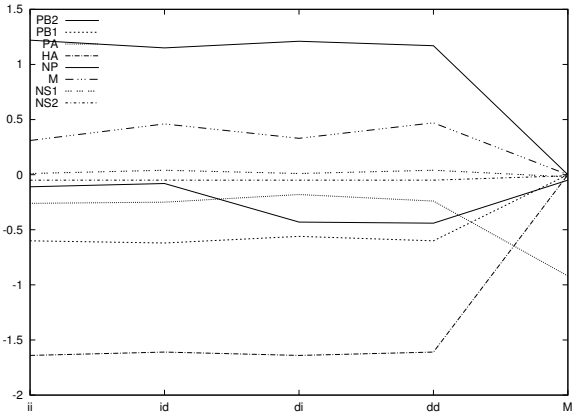

図 7 特徴選択結果を用いたゆらぎで選択された特徴でのクラス分類精度

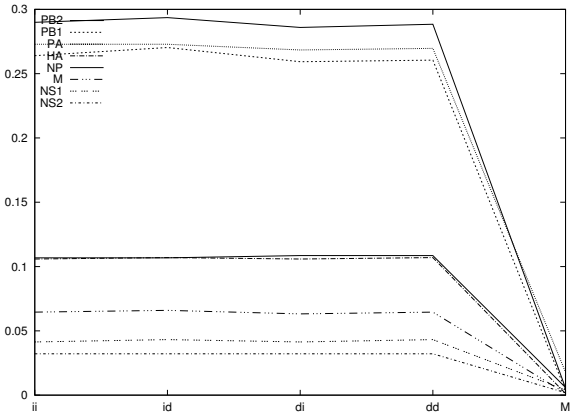

図 8 特徴選択結果を用いたゆらぎで選択された結果に含まれる新規特徴の割合

5. まとめ

本稿では, まず, 対称不確実性のゆらぎを含んだ整列について定義した．次に, 特徴選択手法の CWC アルゴリズムで用いられている対称不確実性によるソートに対し, 定義したゆらぎを含んだ整列を用いて特徴選択を行うアルゴリズムを提案した．最後に, NCBI より提供されているインフルエンザウイルスの RNA データに適用し, 機械学習の精度に影響しない別の特徴集合を得ることができた．結果から定義したゆらぎについて, それぞれ新規特徴の取得, 特徴数の増減, 機械学習精度の増減の 3 点から評価する．部分降順ゆらぎでは, 選択された個々の特徴が持つクラスへの関連性が低下するが, 新規特徴を多く得られる．部分移動ゆらぎでは, データセットごとに閾値に対する影響が異なり, データの特性について検証する場合などに利用できる．未選択特徴優先ゆらぎは, 選択された個々の特徴が持つクラスへの関連性を低下させずに多くの新規特徴を得ることができる．先頭特徴優先ゆらぎは, 選択された特徴にあまり変化がなく, 新規特徴の取得にも不向きである．しかし, このことより対称不確実性の昇順で特徴選択した結果得られた特徴集合のうち, 最小の対称不確実性よりも更に対称不確実性が低い特徴は, 対称不確実性の高い特徴との相互作用する確率は低いといえる．

それぞれのゆらぎから得られた特性を利用し, 新規特徴の取得が多く, 特徴数増減が小さく, 関連性が増加するような整列を見つけることは今後の課題である．特に先頭特徴優先ゆらぎの結果から不要な可能性の高い範囲が見えてきたため, 対称不確実性の昇順に並べた特徴列から選択された先頭特徴より前を除外したデータセットで再度別の順番に整列しなおし, 特徴選択を行うことなどが考えられる．

参考文献

[1] K, Shin.: Super-CWC and Super-LCC: Super Fast Feature Selection Algorithms, *Proc. Big Data'15, 61-67*,

Vol.2018-MPS-118 No.25
Vol.2018-BIO-54 No.25
2018/6/15

2015.

[2] K. Shin, D. Fernandes, S. Miyazaki: Consistency Measures for Feature Selection: A Formal Definition, Relative Sensitivitiy Comparison and a Fast Algorithms, *Proc. IJCAI'11, 1491-1497*, 2011.

[3] K. Shin, S. Miyazaki: A Fast And Accurate Feature Selection Algorithm Based On Binary Consistency Measures, Comput. Intel. 32, 646-666, 2016.

[4] NIPS. Neural Information Processing Systems Conference 2003: Featureselection challenge, 2003.

[5] WCCI. IEEE World Congress on Computational Intelligence 2006: Performance prediction challenge, 2006.

[6] L.C. Molina, L. Belanche, A. Nebot, Feature Selection Algorithms: A Survey and Experimental Evaluation, Proc. IEEE Int'l Conf. Data Mining, pp. 306-313, 2002.

[7] V. Kumar and S. Minz, "Feature selection: A literature review," SmartComputing Review, vol. 4, no. 3, pp. 211-229, June 2014

[8] Z. Zhao and H. Liu. Searching for interacting features. In Proceedings of International Joint Conference on Artificial Intelligence, pages 1156-1161, 2007.

[9] M. Dash, H. Liu, "Consistency-based Search in Feature Selection", Artificial Intelligence, vol. 151, issue 1-2, pp.155-176, 2003.

[10] Y. Bao, P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, D. Lipman: The influenza virus resource at the National Center for Biotechnology Information, J. Virol. **82**, 596–601. Also available at: http://www.ncbi.nlm.gov/genomes/FLU/.