特徴値選択における正規化相互情報量と分類正解率の関係分析

堂本　凌祐† 　申 　吉浩†† 　大島　裕明†

† 兵庫県立大学　応用情報科学研究科　〒 650–0047 兵庫県神戸市中央区港島南町 7–1–28
†† 学習院大学　〒 171–8588 東京都豊島区目白 1–5–1
E-mail: †{aa18e504,ohshima}@ai.u-hyogo.ac.jp, ††yoshihiro.shin@gakushuin.ac.jp

あらまし　本論文ではフィルター型アプローチの特徴値選択アルゴリズムの評価に正規化相互情報量の使用を提案する．特徴値選択とはラベルに関係している特徴値を抽出する手法であり,「ラベルを必要十分に説明している特徴値の抽出（モデル構築）」と「機械学習の性能と効率の改善」の二つの目的を持っている．フィルター型アプローチは，分類器を使用せずにデータセットだけを見て特徴値を選択する手法であるが，従来のフィルター型アプローチの特徴値選択アルゴリズムが特徴値選択の二つの目的を満たしているかどうかを評価するのに複数の分類器が使用されている．しかし，分類器がデータを分類する際に重視する特徴値は分類器によって異なっている．そのため，フィルター型アプローチが抽出した特徴値の組み合わせ F が特定の分類器の性能を改善したからといって，F がラベル C を必要十分に説明している保証もなく，また，他の分類器の性能と効率を改善する保証もない．従って，分類器では特徴値選択の二つの目的を考慮してフィルター型アプローチの特徴値選択アルゴリズムを評価することができない．そこで，本論文では，特徴値選択の二つの目的を考慮してフィルター型アプローチの特徴値選択アルゴリズムを評価することが出来る指標を見つけることを問題とする．この問題に対して，本論文では正規化相互情報量の使用を提案する．正規化相互情報量は,「ラベルを必要十分に説明している特徴値の抽出」を考慮してフィルター型アプローチの特徴値選択アルゴリズムを評価することができると考えられる．しかし，正規化相互情報量が「機械学習の性能と効率の改善」を考慮して評価できるのか不明であるため，本論文では，正規化相互情報量と分類正解率，特徴値数の関係を調査した．

キーワード　特徴値選択アルゴリズム評価，正規化相互情報量

1 はじめに

近年，機械学習は高次元データセットを使用する機会が増えている．この高次元データセットをそのまま使用すると，以下の問題が生じる可能性がある．

- 分類器に入力する情報量が多すぎるため，学習の効率が悪くなる
- ラベルに関係しない情報が学習の妨げを行い，分類器の性能を下げてしまう

上記の問題を解決する手法の一つとして特徴選択，特徴値選択がある．特徴選択，特徴値選択とはラベルに関係している特徴，特徴値を抽出する手法で二つの目的を持っている．

- ラベルを必要十分に説明している特徴，特徴値の抽出（モデル構築）
- 機械学習の性能と効率の改善

しばしば「機械学習の性能と効率の改善」だけに注目されがちであるが，特徴選択，特徴値選択はデータセットを理解する手助けをすることも役割の一つであるため,「ラベルを必要十分に説明している特徴，特徴値の抽出」も重要な特徴選択，特徴値選択の目的の一つである．

特徴選択と特徴値選択の違いは，特徴選択は特徴毎に抽出する手法であるのに対して，特徴値選択は，特徴の代わりに特徴値を抽出する．特徴値選択は特徴だけでなくどの値がラベルを特徴付けているのかを示すため，特徴選択よりも詳細なモデル構築を可能とするので，本論文では下記で説明する問題に対する提案手法の効果を特徴値選択で観察した．しかし，特徴値選択は「One-Hot Encoding」によって，特徴値を二値の値として持つ特徴と見做すことができるので，以下では特徴値と特徴を区別しない．

現在の特徴選択アルゴリズムでは以下の二つを実施することができる．

- 冗長，有効でない特徴の選択を避けること
- 影響しあっている特徴を発見すること

「冗長，有効でない」,「影響しあっている」を表 1 を用いて説明する．表 1 は「当選」,「旅行」,「0 円」,「連絡」,「様」の五種類のワードがデータに含まれているかどうかを表現した特徴ベクトルと,「スパムメール」と「非スパムメール」の二種類の値からなる「ラベル」で構成されているデータセットである．表 1 の「0 円」と「連絡」に注目すると，各特徴の値は各データに対して同じ値をとっている．このような同じ値をとる特徴は一つで十分であり，もし同じ値をとる複数の特徴を分類器に入力すると，分類器の性能は改善されずに分類器の学習効率が悪くなるという問題が発生する．よって，同じ値をとる特徴は一つだけ選択されるべきである．例えば,「0 円」を必要な特徴とすると,「連絡」は不必要な特徴となる．この「連絡」のような特徴を「冗長な特徴」と呼び，特徴ベクトルから削除する．次に，表 1 の「様」に注目すると，全てのデータに「様」が含まれて

表 1　特徴選択の例

当選	旅行	0 円	連絡	様	ラベル
0	1	1	1	1	スパムメール
0	1	1	1	1	スパムメール
0	1	0	0	1	非スパムメール
0	0	1	1	1	非スパムメール
0	0	0	0	1	非スパムメール
1	0	0	0	1	スパムメール
1	1	1	1	1	スパムメール

おり,「様」はラベルに全く関係していない.「有効でない特徴」とは,「様」のような全くラベルに関係しない特徴を指す.「有効でない特徴」は分類器の学習を妨げ，分類器の性能を下げる原因となるため，このような特徴は削除することが望ましい．

表 1 は「当選」というワードが含まれているデータは必ず「スパムメール」となる．しかし,「当選」というワードが含まれていないデータは「スパムメール」と「非スパムメール」の二種類が存在する.「当選」というワードが含まれていない時は,「旅行」と「0 円」の二つのワードが含まれているときは「スパムメール」となり，それ以外の時は「非スパムメール」となる．よって,「ラベル」は「当選」,「旅行」,「0 円」の三つの特徴を組み合わせることで十分に説明することができる．しかし,「当選」,「旅行」,「0 円」は，単体では十分にラベルを説明することができない.「影響しあっている特徴」とは，この例のような単体ではラベルを十分に説明することができないが，複数の特徴を組み合わせることで飛躍的にラベルを説明することができる特徴達を指し，ラベルを必要十分に説明するために不可欠である．

特徴選択には三つのアプローチが存在する．ラッパー型アプローチ，埋め込み型アプローチ，フィルター型アプローチの三つである．ラッパー型アプローチ [2], [3] は，特定の分類器の正解率などが高くなるように特徴を抽出する手法である．埋め込み型アプローチは特徴選択アルゴリズムが分類器の学習に組み込まれている手法であり，決定木などが埋め込み型アプローチの例である．フィルター型アプローチは，データセットだけを見て特徴を抽出する手法である．本論文ではフィルター型アプローチを対象とする．フィルター型アプローチは分類器を使用せず，データセットだけを見て特徴を抽出する手法であるが，従来のフィルター型アプローチの評価には分類器が使用されている．しかし，分類器が特徴をどのように評価して分類に反映をさせているかは分類器毎の「癖」によって異なる．そのため，フィルター型アプローチが抽出した特徴の組み合わせ F が特定の分類器の性能と効率を改善したからといって，F がラベルを必要十分に説明している保証はない．また，他の分類器の性能と効率を改善する保証もない．よって，分類器では特徴選択の二つの目的を考慮してフィルター型アプローチを評価することはできない．そこで，本論文では，フィルター型アプローチを特徴選択の二つの目的を考慮して評価できる指標を見つけることを問題とし，この問題の解決策を考える．

2 節では関連研究について説明する．3 節では本論文で取り扱う問題とその問題に対する提案アプローチについて説明する．4 節では提案アプローチの評価を行い，5 節で本論文のまとめを行う．

2　関 連 研 究

2.1　特徴選択構造

[4] では，特徴選択は定められた「探索方向」に基づいて「探索手段」が出力する特徴集合を「探索指標」によって評価し特徴の採否を判断するというプロセスを繰り返すというモデルに基づいて,「探索方向」,「探索手段」,「探索指標」を特徴選択アルゴリズムを形成する三つの軸と定めている．

探索手段とは，特徴の全ての組み合わせ空間中（冪集合）から，どの特徴集合を評価するのか決定する手段で，例えば全領域探索，ソート探索，ランダム探索が存在する．全領域探索は，全ての特徴の組み合わせを評価対象とし，求められる限り，定められた順序で全ての特徴集合を出力する．しかし，特徴の数を n に対して冪集合のサイズは 2^n になってしまうため，現実的な方法では探索空間を冪集合の部分空間に制限する．例えば，ソート探索では，特徴を特定の順序でソートし，常に順位の高い特徴が優先して含まれるように特徴集合を選択することで探索空間を狭めている．ランダム探索は，局所解に捕まることを防ぐために，ランダムに特徴集合を選択する手法である．

探索方向とは，特徴を選択する向きを意味し，特徴集合のサイズが増加する方向で探索を行う追加型，特徴集合のサイズが減少する方向で探索を行う削除型，追加型と削除型を組み合わせる結合型などが存在する．

探索指標は探索手段が出力した特徴集合を評価して，次に探索手段が出力するべき特徴集合を指定する．情報量，一貫性指標，分類器の指標（正解率や F 値など），データ間の距離などが探索指標として利用されている．一貫性指標とはデータセットに含まれているデータとラベルとの不適合さを示したもので，ベイズリスクなどが一貫性指標に含まれる [6]．また，本論文で取り扱う問題はフィルター型アプローチを正しく評価することが出来る指標を見つけることであり，最適な「探索指標」を見つける問題には取り組まない．

2.2　探 索 指 標

2.2.1　ベイズリスク

ベイズリスクはデータセット S に含まれている矛盾しているデータの割合を示す．また，S は m 個の特徴を含むものとし，$p(f_1 = x_1, ..., f_m = x_m, C = c)$ を S から導かれる経験的確率分布とする．すなわち，$p(f_1 = x_1, ..., f_m = x_m, C = c)$ は特徴が $f_i = x_i$，クラスラベルが $C = c$ となる S 中のインスタンスの m に対する比率と定義する．

$$ICR(S) = 1 - \sum_{(x_1,.,x_m)\in S} max_c p(c|f_1 = x_1, ., f_m = x_m) \tag{1}$$

矛盾しているデータとは特徴の値が同じであるのに，ラベルが異なっているデータを指す．例えば，二つデータ $d_1 = [1, 0, 1, 0, a]$

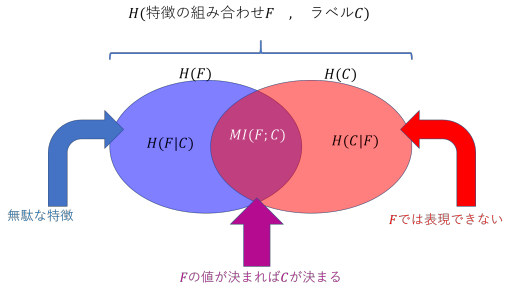

図 1　$H(F)$ と $H(C)$

と $d_2 = [1, 0, 1, 0, b]$（1 と 0 は特徴の値を，a と b はラベルを意味する）は矛盾しているデータである．またベイズリスクは $H(C|F)$ と関係していることが分かっている [7]．つまり，$H(C|F)$ の値が分かれば，おおよそのベイズリスクの値を予測することが可能であり，逆も可能である．つまり，「$H(C|F) = 0 \Leftrightarrow I(F;C) = H(C) \Leftrightarrow ICR(S_F) = 0$」である．$H()$ は平均情報量（エントロピー）を指し，S_F は特徴の組み合わせが F でのデータセットを表している．

2.2.2　相互情報量（Mutual Information）

特徴の組み合わせ F とラベル C の相互情報量について説明する．相互情報量 $MI(F;C)$ は F の値が決まれば C を決定することができる情報量を意味し，次式で求められる．

$$\begin{aligned} MI(F;C) &= H(F) + H(C) - H(F,C) \\ &= \sum_{f \in F, c \in C} p(f,c) \log_2 \frac{p(f,c)}{p(f)p(c)} \end{aligned} \tag{2}$$

2.2.3　正規化相互情報量（Normalized Mutual Information）

正規化相互情報量を説明するために，図 1 の説明を行う．図 1 は特徴の組み合わせ F のエントロピー $H(F)$ とラベル C のエントロピー $H(C)$ を示している．図 1 の色がついた各領域にはそれぞれ意味があり，青色の領域は F に含まれる有効でない特徴の情報量 $H(F|C)$，紫の領域は特徴の値が決まればラベルを決定することができる情報量 $MI(F;C)$，赤色の領域は F では表現することが出来ない C の情報量 $H(C|F)$ を示している．正規化相互情報量 NMI は，青色，紫色，赤色の全ての領域を考慮して F を評価することができる指標であり，次式で求められる．

$$NMI(F,C) = \frac{2MI(F;C)}{H(F) + H(C)} \tag{3}$$

正規化相互情報量は $\frac{MI(F;C)}{H(F)}$ と $\frac{MI(F;C)}{H(C)}$ の調和平均で求められるため，正規化相互情報量の値が高いほど，特徴の組み合わせはラベルと相関しており，かつ，有効でない特徴が含まれていないと考えることができる．

2.3　特徴選択アルゴリズムの紹介

ここでは探索指標の種類ごとに特徴選択アルゴリズムの一部を紹介する．まず初めに，一貫性指標を用いたアルゴリズムを紹介する．

一つ目は INTERACT [10] である．INTERACT の探索手段はソート探索，探索方向は削除型，探索指標は cc（consistency contribution) である．特徴の組み合わせを F'，F' でのデータセットを $S_{F'}$，i 番目の特徴を f_i とすると，cc は次式で表される．

$$cc(F', f_i) = ICR(S_{F' \setminus f_i}) - ICR(S_{F'}) \tag{4}$$

INTERACT は全ての特徴を選択している状態 F' からスタートする．INTERACT は最初に各特徴 f とラベル C 間の正規化相互情報量 $NMI(f,C)$ が降順になるように F' を並び替える．そして，正規化相互情報量の値が高い特徴から $cc(F', f_i)$ を計算し，$cc(F', f_i)$ が閾値を超え無ければ，f_i は必要のない特徴として F' から削除される．

二つ目は Super CWC [6] である．Super CWC の探索手段はソート探索，探索方向は削除型，探索指標はベイズリスクである．Super CWC は全ての特徴を選択している状態 F' からスタートする．Super CWC は最初に特徴の組み合わせ F' でのデータセット $S_{F'}$ のベイズリスク $ICR(S_{F'})$ の値を計算する．もし，$ICR(S_{F'})$ が 0 でなければ，$ICR(S_{F'})$ が 0 になるようにデータセットからデータを削除する．$ICR(S_{F'})$ が 0 になると，各特徴 f とラベル C との間の正規化相互情報量 $NMI(f,C)$ が昇順になるように特徴の組み合わせ F' を並び替える．そして，二分探索を用いて $ICR(S_{F' \setminus \{f_{t+2}, .., f_i\}}) = 0$ となる最大の i を探索し，t を最後に削除した特徴の添え字として $\{f_{t+2}, ..., f_i\}$ を F' から全て削除する．

次は情報量を用いたアルゴリズムを紹介する．一つ目は mRMR [5] である．mRMR の探索手段はソート探索，探索方向は追加型，探索指標は分類器の誤り分類率と $mRMR$ である．特徴の数が m 個の特徴の組み合わせを $F_m = \{f_1, .., f_m\}$，ラベルを C とすると，$mRMR$ は次式で表される．

$$mRMR(F_m) = \sum_{i=1}^{m} \left(MI(f_i;C) - \frac{1}{m} \sum_{j \neq i} MI(f_i;f_j) \right) \tag{5}$$

mRMR は選択している特徴の数が 0 個の状態からスタートする．一回目の選択では各特徴の $mRMR$ を計算し，最も大きい値を示した特徴を選択する．二回目以降の選択では，選択回数を t とすると，$t-1$ 回目までに選択されていない各特徴と F_{t-1} との $mRMR$ を計算し，最も大きい値を示した特徴を F_{t-1} に追加し，そのときの特徴の組み合わせを F_t とする．上記の操作を特徴の数がユーザによって指定された数 n になるまで続ける．すると，特徴の組み合わせは F_1 から F_n の n 個作成することが出来る．F_1 から F_n のうち，分類器の誤り分類率が最も低い特徴の組み合わせを最適な組み合わせとして出力する．

二つ目は FCBF [9] である．FCBF の探索手段はソート探索，探索方向は削除型，探索指標は正規化相互情報量である．FCBF は最初に各特徴 f とラベル C 間の正規化相互情報量 $NMI(f,C)$ を計算し，閾値を超えていない特徴は全て削除し，残った特徴を正規化相互情報量が降順になるように並び替える．そして，残った特徴の組み合わせ $F' = \{f_1, ..., f_k\}$ から更に不必要な特徴を削除する．削除の方法は F' から二つの特徴 f_i と f_j

($i < j$) を取り出し，二つの特徴が $NMI(f_i, f_j) > NMI(f_j, c)$ の関係を満たすならば，f_j は必要のない特徴として F' から除去する．上記の操作を全ての特徴に対して行い，残った特徴を最適な特徴の組み合わせとして出力する．

最後にデータ間の距離を使用したアルゴリズムを紹介する．Relif [1] の探索方法はランダム探索，探索方向は削除型，探索指標はユーザが指定する閾値 t とデータ間の距離である．データ数が m 個，各データにラベル c がついているデータセットを D, 一つのデータは n 個の特徴から構成されている $F = \{f_1, ..., f_n\}$ とする．Relif は各特徴に重み w をつけ $W = \{w_1, w_2, ..., w_n\}$, この重みに従って特徴を削除していく．全ての w は 0 の状態からスタートする $W = \{0, 0, ..., 0\}$．Relif は最初に D からランダムに一つのデータ $X = \{x_1, ..., x_n\}$ を選択する．さらに，各ラベルのデータの中で X に最も距離が近いデータをそれぞれ抽出する（ここではラベルは a と b の二種類と考え，X に近いデータをそれぞれ $Z_a = \{z_{a1}, ..., z_{an}\}$，$Z_b = \{z_{b1}, ..., z_{bn}\}$ とする）．距離 *diff* はバイナリデータではマンハッタン距離が，バイナリデータでなければユークリッド距離が使用される．各特徴の重み w_i は，ランダムに抽出した X と X に近い二種類のデータ Z_a，Z_b を用いて次式で更新される．

X のラベルが a のとき

$$w_i = w_i - diff(x_i, z_{ai})^2 + diff(x_i, z_{bi})^2 \qquad (6)$$

X のラベルが b のとき

$$w_i = w_i - diff(x_i, z_{bi})^2 + diff(x_i, z_{ai})^2 \qquad (7)$$

上記の操作を m 回行い，その後，各重みを m で割って平均を出す $W = \{\frac{w_1}{m}, ..., \frac{w_n}{m}\}$．最後に各特徴の重み w_i がユーザによって指定された閾値 t を超えていなければ，特徴 f_i は必要のない特徴として削除し，残った特徴の組み合わせが出力される．

3 問題定義と提案アプローチ

3.1 問題定義

特徴選択は「ラベルを必要十分に説明している特徴の抽出（モデル構築）」と「機械学習の性能と効率の改善」の二つの目的を持っている．特徴選択手法の一つであるフィルター型アプローチは分類器を使用せず，データセットだけを見て特徴を抽出する手法であるため，特徴選択の目的の一つである「ラベルを必要十分に説明している特徴の抽出」を考慮していると考えられるが，従来ではフィルター型アプローチの特徴選択アルゴリズムが特徴選択の二つの目的を考慮しているかの最終的な判断には分類器が使用されている．しかし，各分類器はそれぞれ固有の癖を持っており，データを分類する際に重視する特徴が分類器によって異なっている．そのため，同じデータセットを Support Vector Machine と Naive Bayes の二つの分類器に入力したとき，Support Vector Machine では高い正解率を示し，Naive Bayes では低い正解率を示すということがあり得る．よって，フィルター型アプローチの特徴選択アルゴリズムが抽出した特徴の組み合わせ F が特定の分類器の性能と効率を改善したからといって，F がラベルを必要十分に説明している保証はなく，また，他の分類器の性能と効率を改善する保証もない．つまり，分類器では特徴選択の二つの目的を考慮してフィルター型アプローチの特徴選択アルゴリズムを評価することが出来ない．そこで，本論文ではフィルター型アプローチの特徴選択アルゴリズムを特徴選択の二つの目的を考慮して評価することが出来る指標を見つけることを問題とし，この問題を解く．

3.2 特徴選択アルゴリズム評価指標の条件

一般的に理想的な特徴の組み合わせは以下の二つの条件を満たしている．

- 少ない特徴でラベルを表現することができる
- 特徴の組み合わせとラベルとの間に強い相関がある

「少ない特徴でラベルを表現すること」ができれば，特徴の組み合わせに冗長な特徴や有効でない特徴が含まれていないことを意味するため，分類器の性能と効率の改善が可能である．また,「少ない特徴でラベルを表現することができる」と「特徴の組み合わせとラベルとの間に強い相関がある」を満たせば，その特徴の組み合わせがラベルを必要十分に説明することができていることを意味する．以上二つの条件を踏まえて，特徴選択アルゴリズムの評価指標の条件について考える．一般的に，特徴選択は探索指標の値が基準値を超える，または，最適値を示す特徴の組み合わせのうち，最も特徴の数が少ない組み合わせを選択する．よって，特徴選択アルゴリズムの評価指標が冗長な特徴を評価する必要性は低い．以上を踏まえて，特徴選択アルゴリズム評価指標の条件は以下の四つであると考える．

- 有効でない特徴を考慮して評価することができる
- 特徴の組み合わせとラベルとの間の相関を評価することができる
- 分類器の性能と効率の改善を考慮して評価することができる
- 分類器を使用せずに評価することができる

本論文では，上記四つの条件を満たす評価指標について考え，その指標の提案を行う．

3.3 問題に対する提案アプローチ

3.1 節の問題の解決策として，フィルター型アプローチを形成するのに必要な要素の一つである「探索指標」を特徴選択アルゴリズムの評価指標として使用することが考えられる．一般的にフィルター型アプローチの探索指標としては，相互情報量やベイズリスクが使用されている．相互情報量やベイズリスクは「特徴の組み合わせ F が決まればラベル C を決定することができる情報量」や,「F では表現することが出来ない C の情報量」などを考慮して F を評価することができる．しかし,「F に含まれる有効でない特徴」を考慮して評価することは出来ないため，3.2 節の一つ目の条件を考慮していない．そこで，我々は特徴選択アルゴリズムの評価指標に正規化相互情報量の使用を提案する．その理由は，正規化相互情報量は「F に含まれる有効でない特徴の情報量」,「F が決まれば C を決定することができる情報量」,「F では表現することができない C の情報量」

表 2　monks データセット

データ名	データ数	特徴数	ラベル数
monks1	432	17	2
monks2	432	17	2
monks3	432	17	2

を考慮して F を評価することが可能であるためである．三つの情報量を考慮して F を評価するため，正規化相互情報量は 3.2 節の一つ目，二つ目，四つ目の条件は考慮していると考えられる．つまり，正規化相互情報量は特徴選択の目的の一つである「ラベルを必要十分に説明している特徴の抽出」を相互情報量やベイズリスクよりも考慮してフィルター型アプローチを評価することができる．しかし，正規化相互情報量が特徴選択のもう一つの目的である「機械学習の性能と効率の改善」を考慮してフィルター型アプローチの特徴選択アルゴリズムを評価することができるのか不明である．そこで，本論文では，正規化相互情報量の値と分類正解率や特徴数との関係を分析し，正規化相互情報量がフィルター型アプローチの特徴選択アルゴリズムの評価指標に適しているのか議論する．

4　各指標と分類正解率，特徴数の関係

4.1　実 験 概 要

本論文では悉皆法を用いて正規化相互情報量の値と分類正解率，特徴数の関係を分析する．正規化相互情報量の値の上昇とともに，分類正解率も上昇すれば，正規化相互情報量は分類器の性能を考慮して特徴の組み合わせを評価することができると考えられる．

4.3 節では，正規化相互情報量と特徴数の関係の結果を，4.4 節では正規化相互情報量と分類正解率の関係の結果を記載する．また，4.3 節，4.4 節では，正規化相互情報量の比較対象として相互情報量と分類正解率，特徴数の関係も分析する．

また，本実験の結果の中には正規化相互情報量が最大にも関わらず，分類正解率が最大値よりも低い結果を示すデータセットが存在した．4.5 節では，上記の問題が生じた原因について考察する．

4.2　実 験 環 境

今回の実験は悉皆法を使用して行った．悉皆法を使用しているため，特徴数が n とすると，考える特徴の組み合わせは 2^n となる．そのため，本論文では表 2 に示す特徴数の少ない三種類のデータセット [8] を用いて実験をおこなった．また，本論文では k Nearest Neighbor（kNN）と Naïve Bayes（NB），Support Vector Machine（SVM）の三種類の分類器を使用して正規化相互情報量と分類正解率の関係を分析した．実験はデータの分割数を 3，random state を 7 とした sklearn の StratifiedKFold の交差検証を使用して行った．各分類器のモデル構築に必要なパラメータはグリッドサーチを用いて実験を行った．以下に今回の実験で使用したそれぞれの分類器の設定を記す．

kNN は sklearn.neighbors の KNeighborsClassifier を使用して実験を行った．KNeighborsClassifier の引数である metric は

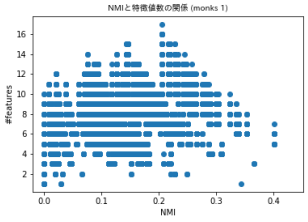

図 2　NMI と特徴数の関係 $(monks1)$

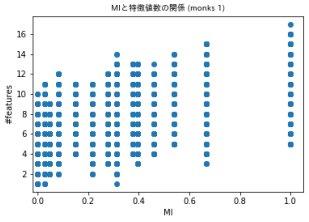

図 3　MI と特徴数の関係 $(monks1)$

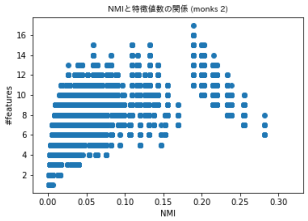

図 4　NMI と特徴数の関係 $(monks2)$

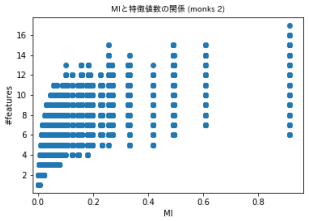

図 5　MI と特徴数の関係 $(monks2)$

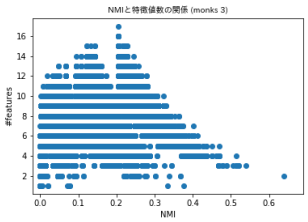

図 6　NMI と特徴数の関係 $(monks3)$

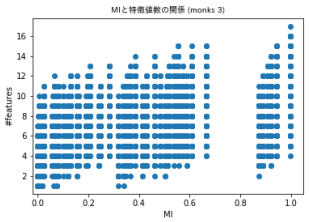

図 7　MI と特徴数の関係 $(monks3)$

minkowski とし，p は 1 に設定した．metric を minkowski，p を 1 に設定することで，データの距離をマンハッタン距離によって計算することが可能になる．また，k のパラメータは 1〜10 とし，グリッドサーチを使用して，それぞれの特徴の組み合わせ毎に最適な k の値を設定した．

NB は sklearn の naive_bayes の BernoulliNB を用いて実験を行った．BernoulliNB の引数である binarize は None と設定した．

SVM は sklearn.svm の SVC を使用して実験を行った．引数である kernel は rbf と設定した．そのため，本実験では元々 SVC で使用するパラメータ C と rbf カーネルで使用するパラメータ gamma が必要になる．それぞれのパラメータの範囲は C では [0.1, 1, 10, 100, 1000, 10000] の値をとり，gamma では [0.1, 1, 10, 20, 50, 70, 100] の値をとるように設定した．そして，それぞれの特徴の組み合わせ毎に最適な C と gamma をグリッドサーチで決定した．

4.3　各指標と特徴数

図 2，図 3 は monks1 での各指標と特徴数の結果を，図 4，図 5 は monks2 での各指標と特徴数の関係を，図 6，図 7 は monks3 での各指標と特徴数の関係を示している．全ての結果において各指標の最大値での特徴数に注目すると，正規化相互情報量は相互情報量と比べて特徴数のふり幅は小さく，また，特徴数の最大値は正規化相互情報量のほうが圧倒的に少ない．

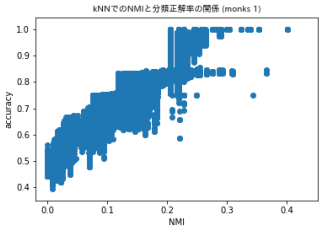

図 8　*kNN*での *NMI*と分類正解率の関係 (*monks1*)

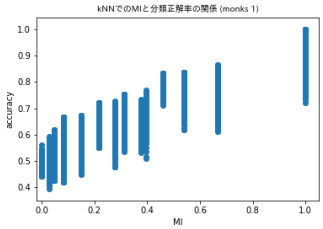

図 9　*kNN*での *MI*と分類正解率の関係 (*monks1*)

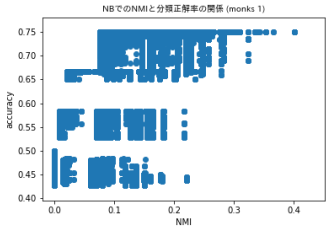

図 14　*NB*での *NMI*と分類正解率の関係 (*monks1*)

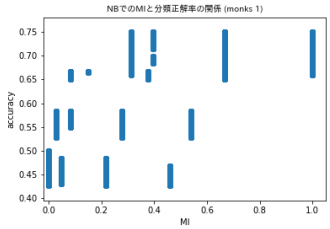

図 15　*NB*での *MI*と分類正解率の関係 (*monks1*)

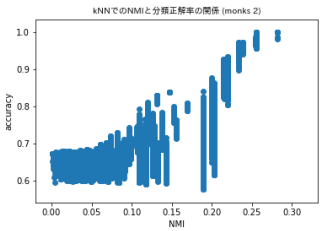

図 10　*kNN*での *NMI*と分類正解率の関係 (*monks2*)

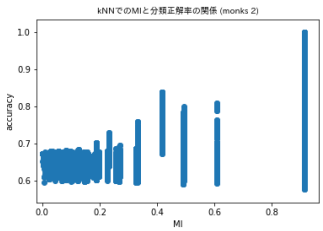

図 11　*kNN*での *MI*と分類正解率の関係 (*monks2*)

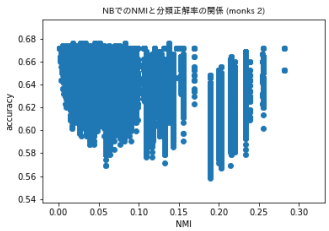

図 16　*NB*での *NMI*と分類正解率の関係 (*monks2*)

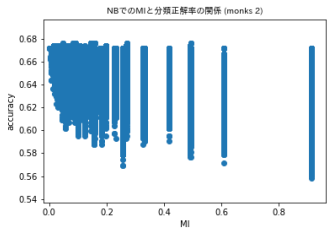

図 17　*NB*での *MI*と分類正解率の関係 (*monks2*)

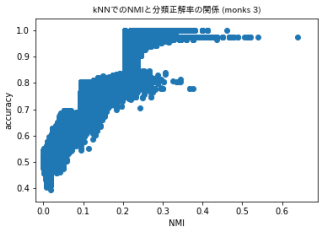

図 12　*kNN*での *NMI*と分類正解率の関係 (*monks3*)

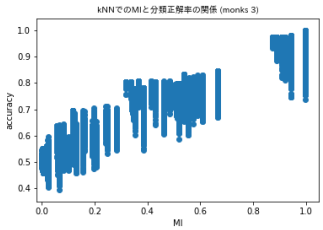

図 13　*kNN*での *MI* と分類正解率の関係 (*monks3*)

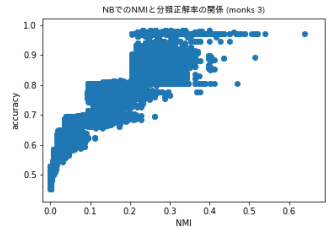

図 18　*NB*での *NMI*と分類正解率の関係 (*monks3*)

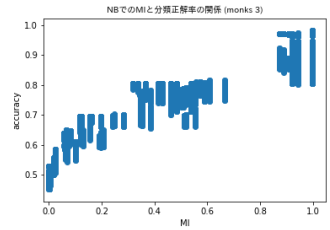

図 19　*NB*での *MI*と分類正解率の関係 (*monks3*)

その理由は，正規化相互情報量は $H(F|C)$ を考慮して評価できるのに対して，相互情報量は $H(F|C)$ を考慮して評価することが出来ないからであると考える．少ない特徴でラベルを表現することができるほど，分類器に入力される情報量は少なくなり機械学習の効率は改善されるので，正規化相互情報量は相互情報量よりも「機械学習の効率の改善」を考慮してフィルター型アプローチの特徴選択アルゴリズムを評価することができると考えられる．

4.4　各指標と分類正解率

4.4.1　kNN の結果の分析

図 8，図 9 は monks1 での各指標と kNN の分類正解率の関係を，図 10，図 11 は monks2 での各指標と kNN の分類正解率の関係を，図 12，図 13 は monks3 での各指標の値と kNN の分類正解率の関係を示している．全ての結果で正規化相互情報量，相互情報量の値が大きくなるほど分類正解率も上昇することを確認することができた．また，各指標が最大値での分類正解率に注目すると，相互情報量では分類正解率に大きな振れ幅があるのに対し，正規化相互情報量では分類正解率に大きな振れ幅は無く，最大値に近い値を示す結果となった．以上より，kNN では相互情報量が最大値を示す特徴の組み合わせの分類正解率が必ず最大を示すとは限らず，対して，正規化相互情報量が最大値を示す特徴の組み合わせの分類正解率は最大値に近い値を示すということが分かった．以上より，kNN では正規化相互情報量は分類正解率と正の相関があり，相互情報量よりもフィルター型アプローチの評価指標として適していることが分かった．しかし，monks3 において正規化相互情報量が最大値を示すにも関わらず，分類正解率が最大値よりも少し低い結果となった．この原因は 4.5 節で考察する．

4.4.2　NB の結果の分析

図 14，図 15 は monks1 での各指標と NB の分類正解率の関係を，図 16，図 17 は monks2 での各指標と NB の分類正解率の関係を，図 18，図 19 は monks3 での各指標と NB の分類正解率の関係を表している．monks1，monks3 では正規化相互情報量，相互情報量の値が高くなるほど分類正解率も上昇することを確認できた．monks2 に限っては，各指標の値が上昇しても分類正解率の最大値はほとんど変わらない結果となった．また，各指標の最大値の分類正解率に着目すると，相互情報量では分類正解率に大きなふり幅があるのに対して，正規化相互情報量では大きなふり幅は確認できなかった．従って，相互情報量が最大値を示す特徴の組み合わせの分類正解率は最大値を示すとは限らず，正規化相互情報量が最大を示す特徴の組み合わせの分類正解率は必ず最大に近い値を示すことが分かった．以上より，NB でも正規化相互情報量が相互情報量よりもフィルター型アプローチの特徴選択アルゴリズムの評価指標として適していることが分かった．しかし，monks3 において正規化相

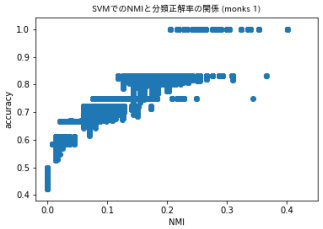

図 20　*SVM*での *NMI*と分類正解率の関係 (*monks1*)

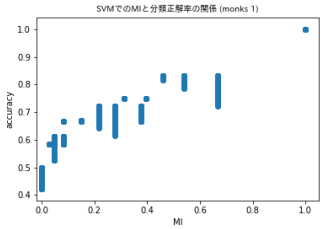

図 21　*SVM*での *MI*と分類正解率の関係 (*monks1*)

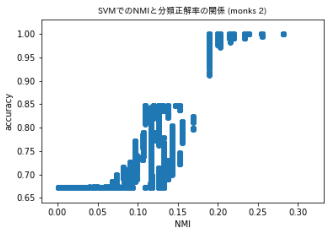

図 22　*SVM*での *NMI*と分類正解率の関係 (*monks2*)

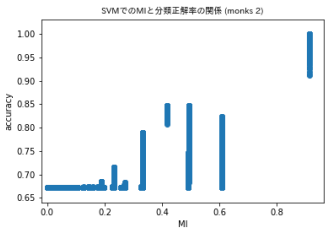

図 23　*SVM*での *MI*と分類正解率の関係 (*monks2*)

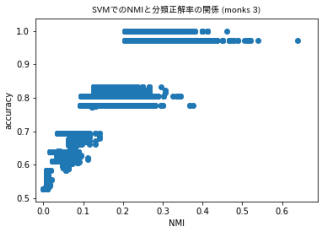

図 24　*SVM*での *NMI*と分類正解率の関係 (*monks3*)

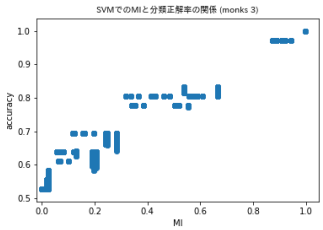

図 25　*SVM*での *MI*と分類正解率の関係 (*monks3*)

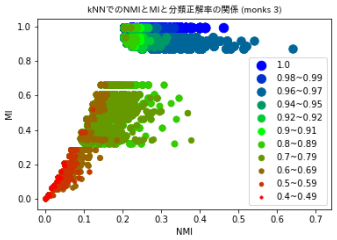

図 26　*kNN*での *NMI*と *MI*と分類正解率の関係 (*monks3*)

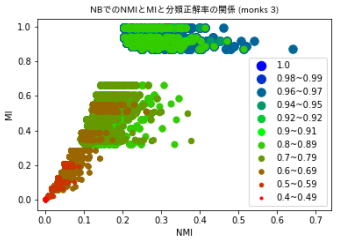

図 27　*NB*での *NMI*と *MI*と分類正解率の関係 (*monks3*)

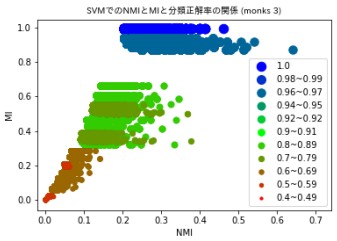

図 28　*SVM*での *NMI*と *MI*と分類正解率の関係 (*monks3*)

互情報量が最大値を示す特徴の組み合わせの分類正解率は最大値よりも少し低い結果となった．この原因は 4.5 節で考察する．

4.4.3　SVM の結果の分析

図 20，図 21 は monks1 での各指標と SVM の分類正解率の関係を，図 22，図 23 は monks2 での各指標と SVM の分類正解率の関係を，図 24，図 25 は monks3 での各指標の値と SVM の分類正解率の関係を示している．結果より，正規化相互情報量，相互情報量の値が大きくなるほど分類正解率も上昇することを確認することができた．また，monks2 の各指標が最大値のときの分類正解率に着目すると，相互情報量では分類正解率にふり幅があるのに対して，正規化相互情報量では分類正解率に大きなふり幅は存在していない．monks2 においては相互情報量が最大値を示す特徴の組み合わせの分類正解率が最大値になるとは限らないのに対して，正規化相互情報量では最大値を示せば分類正解率は最大に近い値を示すことを確認することができた．よって，monks2 においては正規化相互情報量が相互情報量よりも特徴選択アルゴリズムの評価指標として適していることが分かった．しかし，monks1，monks3 においては正規化相互情報量が特別に相互情報量よりも優れていることを確認することが出来なかった．その理由は，SVM が有効でない特徴の影響を受けずに正しく分類することができたためであると考える．monks3 では，正規化相互情報量が最大値を示す特徴の組み合わせの分類正解率は最大値よりも少し低い結果となった．この原因は 4.5 節で考察する．

4.5　monks3 で *NMI* が最大値での分類正解率が最大値にならない原因

図 26，図 27，図 28 は monks3 での正規化相互情報量と相互情報量の関係を示している．また，各プロットの色は各分類器の分類正解率に基づいて色分けを行っている．

monks3 では kNN や NB，SVM において正規化相互情報量が最大値を示す特徴の組み合わせの分類正解率は最大値よりも少し低い結果となった．この原因は，図 26，図 27，図 28 より，正規化相互情報量が最大値のときの相互情報量が最大値よりも少し低いためであると考えられる．正規化相互情報量が最大値であるのに相互情報量が最大値にならなかった原因は，相互情報量が最大値になるために必要な特徴に有効ではない特徴の情報量 $H(F|C)$ が多量に含まれていたためであると考えられる．しかし，正規化相互情報量が最大値での分類正解率と monks3 での最大分類正解率との差は小さい．よって，本実験では monks3 において正規化相互情報量が最大にも関わらず，分類正解率が最大値とならなかったことに重大な問題があると考えず，これにより正規化相互情報量が相互情報量よりも特徴選択アルゴリズムの評価指標として適していないとは考えない．

5 まとめ

本論文ではフィルター型アプローチの特徴選択アルゴリズムの評価指標に正規化相互情報量の使用を提案した．正規化相互情報量は特徴選択の目的である「ラベルを必要十分に説明している特徴の抽出」を考慮してフィルター型アプローチの特徴選択アルゴリズムを評価することができると考えられる．しかし，正規化相互情報量が「機械学習の性能と効率の改善」を考慮して評価できるのか不明であった．そこで，本論文では正規化相互情報量と分類正解率，特徴数の関係を調査した．

4.4 節の結果より，kNN，SVM では全てのデータセットで，NB では monks2 を除くデータセットで正規化相互情報量が上昇すれば分類正解率も上昇することを確認できた．また，相互情報量でも値が上昇すれば分類正解率も上昇する結果となった．しかし，各指標が最大値のときの分類正解率に着目すると，相互情報量では分類正解率にふり幅があるのに対して，正規化相互情報量の分類正解率には大きなふり幅は確認できなかった．よって，正規化相互情報量は「機械学習の性能の改善」を考慮して評価することが出来るのに対して，相互情報量は「機械学習の性能の改善」を考慮して評価することが出来ないことが分かった．

4.3 節でも正規化相互情報量は相互情報量よりもフィルター型アプローチの特徴選択アルゴリズムの評価指標として適していることを確認することが出来た．各指標が最大値での特徴数に着目すると，相互情報量では特徴数に大きなふり幅があるのに対して，正規化相互情報量のふり幅は小さかった．また，各指標が最大での最大特徴数にも着目すると，正規化相互情報量の最大特徴数は相互情報量の最大特徴数よりも圧倒的に少ない．以上より，正規化相互情報量は相互情報量よりも「分類器の効率の改善」を考慮してフィルター型アプローチの特徴選択アルゴリズムを評価することができることを確認することができた．

以上より正規化相互情報量は特徴選択の二つの目的を考慮して特徴の組み合わせを評価することができるため，フィルター型アプローチの特徴選択アルゴリズムの評価指標として適していると考えられる．

謝　辞

本研究の一部は JSPS 科学研究費助成事業 JP16H02906，JP17H00762，JP18H03243 による助成を受けたものです．ここに記して謝意を表します．

文　献

[1] K. Kira and L. A. Rendell. A practical approach to feature selection. In *Proceedings of the 9th International Workshop on Machine Learning*, pages 249–256, 1992.

[2] H. Liu and R. Setiono. Feature selection and classification - a probabilistic wrapper approach. In *Proceedings of the 9th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, pages 419–424, 1996.

[3] S. Maldonado and R. Weber. A wrapper method for feature selection using support vector machines. *Information Sciences*, 179(13):2208–2217, 2009.

[4] L. C. Molina, L. Belanche, and A. Nebot. Feature selection algorithms: A survey and experimental evaluation. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, pages 306–313, 2002.

[5] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[6] K. Shin, T. Kuboyama, T. Hashimoto, and D. Shepard. Super-CWC and Super-LCC: Super fast feature selection algorithms. In *Proceedings of the 2015 IEEE International Conference on Big Data*, pages 1–7, 2015.

[7] K. Shin and X. M. Xu. Consistency-based feature selection. In *Proceedings of the 13th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 342–350, 2009.

[8] S. B. Thrun, J. Bala, E. Bloedorn, I. Bratko, B. Cestnik, J. Cheng, K. D. Jong, S. Džeroski, S. E. Fahlman, D. Fisher, R. Hamann, K. Kaufman, S. Keller, I. Kononenko, J. Kreuziger, R. S. Michalski, T. Mitchell, P. Pachowicz, Y. Reich, H. Vafaie, W. V. D. Welde, W. Wenzel, J. Wnek, and J. Zhang. The MONK's problems: A performance comparison of different learning algorithms. Technical report, Carnegie Mellon University, Computer Science Department, 1991.

[9] L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th International Conference on Machine Learning*, pages 856–863, 2003.

[10] Z. Zhao and H. Liu. Searching for interacting features. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 1156–1161, 2007.