ルールベース機械学習のための離散化手法

Discretization Method for Rule-Based Machine Learning

加藤 孝史[1]　岩下 洋哲[2]　後藤 啓介[2]
高木 拓也[2]　鈴木 浩史[2]　大堀 耕太郎[2]

Takashi Kato[1], Hiroaki Iwashita[2], Keisuke Goto[2],
Takuya Takagi[2], Hirofumi Suzuki[2], and Kotaro Ohori[2]

[1] 富士通九州ネットワークテクノロジーズ株式会社
[1] Fujitsu Kyushu Network Technologies Limited
[2] 株式会社富士通研究所
[2] Fujitsu Laboratories Ltd.

Abstract: ビジネスへの AI 利活用の取組みにおいて, AI 判断の根拠を説明できるホワイトボックスの機械学習手法に注目が集まっている. たとえば, ルールベースの機械学習手法は人間が理解しやすいモデルとして知られている. これらは多数のルールを考慮することで精度を高めることができる. しかし, ルールベースの機械学習はその特徴量として連続値を扱う際には, 前処理として離散化してルールの基準を作成する必要があるため, 前処理の方法によって精度が大きく左右される. 特に, 複数の特徴量から形成される高次のルールを扱う予測モデルを構築する場合には, ある 1 つの特徴量を上手く離散化しても, 他の特徴量との関係によって, 最適な離散化となっていない可能性が高い. 本論文では, ルールベースの機械学習において, 最適な離散化を行うための新たな手法を提案する.

1 はじめに

近年, ビジネスへの AI 利活用の取組みにおいて, AI 判断の根拠を説明できるホワイトボックスの機械学習手法に注目が集まっている[1][2][3]. たとえば, ルールベースの機械学習手法は人間が理解しやすいモデルとして知られている. 顕在パターン列挙の技術は, 機械学習の重みづけと組み合わせることでルールベースの機械学習手法に応用することができる[4][5]. この手法では, データ項目の組合せを網羅的に検証して多数のルールを考慮することにより, 高い精度を得られることが確認されている[5].

しかし, ルールベースの機械学習はその特徴量としてバイナリ変数を取り扱うため, 連続値を扱う際には, 前処理として離散化を行う必要がある. 離散化とは, 定量的データを定性的データに変換するプロセスである. この離散化の方法によって精度が大きく左右される. 機械学習における離散化についてこれまで多くの研究成果が発表されている[6][7]. 多くの離散化手法は 1 つの特徴量に着目して実施するものである. しかし, 複数の特徴量から形成される高次のルールを扱う予測モデルを構築する場合には, ある 1 つの特徴量を上手く離散化しても, 他の特徴量との関係によって, 最適な離散化となっていない可能性が高い. そのため, 高次のルールを扱う場合でも高精度を達成する離散化手法を確立する必要がある.

本論文では, 顕在パターンマイニングと既存の離散化手法の拡張を組み合わせた新しい離散化手法を提案する. 具体的には既存の離散化手法を複数の特徴量の関係性を考慮するように拡張し, それを適用する特徴量の組合せを顕在パターンマイニングで取得する. これにより分類精度が向上することを検証する.

2 関連研究

本提案手法の基本となる単変数の離散化手法について説明する.

2.1 MDLP Discretization

Minimum Data Length Principle Discretization[8]は情報エントロピーを最小化する集合を二分する閾値(カットポイント)を再帰的に探索する教師ありの離散化手法である. また, 最小記述長原理に則りカットを実施するかを決定する.

図 1 に示すように，あるソートされたラベル付きの特徴量で構成されたサイズ$size_S$の集合Sにおいて，あるカットポイントcutで分割された集合AとそのAのラベルの割合$P(A)$，集合のサイズ$size_A$，集合Bも同様にラベルの割合$P(B)$，サイズ$size_B$としたとき，集合Aと集合Bのエントロピー$H(A)$，$H(B)$と，そのカットによるエントロピー$H(cut)$は以下の(1), (2), (3)式により計算できる．

$$H(A) = -P(A)\log_2 P(A) \quad (1)$$

$$H(B) = -P(B)\log_2 P(B) \quad (2)$$

$$H(cut) = \frac{size_A}{size_S}H(A) + \frac{size_B}{size_S}H(B) \quad (3)$$

この$H(cut)$が最小となるカットポイントを，隣接する各要素の中間であるカットポイント候補から抽出する．$S = \{v_1, v_2, v_3, \cdots, v_n\}$であったとき，カットポイント候補は$(v_1 + v_2)/2$, $(v_2 + v_3)/2$, …, $(v_{n-1} + v_n)/2$となる．

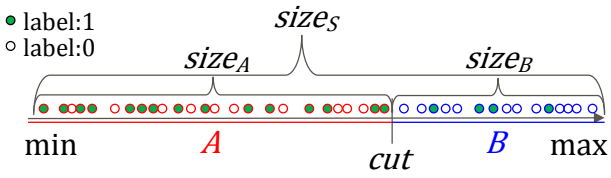

図 1 :あるカットポイントを選択した場合の図

また，k_SをSに出現するクラス数，同様にk_Aとk_BをAとBに出現するクラス数としたときに，$gain$とΔは(4), (5)式と定義する．

$$gain = H(S) - H(cut) \quad (4)$$

$$\Delta = log_2(3^{ks} - 2) - \left(k_S H(S) - K_A H(A) - k_B H(B)\right) \quad (5)$$

カットは以下の(6)式を満たすとき実施する．

$$gain > \frac{\log_2(size_S - 1)}{size_S} + \frac{\Delta}{size_S} \quad (6)$$

このカットポイント探索をカットポイントが抽出されなくなるまで再帰的に実行することで複数のカットポイントを抽出することが可能である．

しかし，この手法のような単変数に対する離散化の欠点として，複数変数間の関係性を考慮した離散化が行われない場合がある．図 2 に離散化が上手く実施できない例を示す．

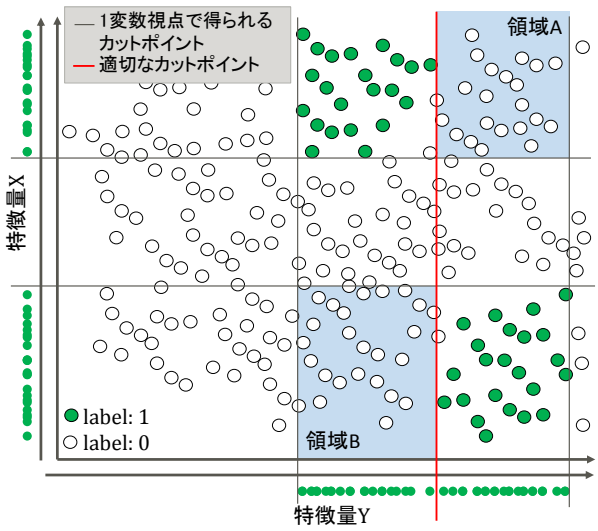

図 2 :複数変数での離散化失敗の例

特徴量 X,Y があり，それぞれにおいて離散化を実施すると黒線のカットポイントは得られるが，赤線のカットポイントは両特徴量を加味して初めて得られるカットポイントである．この手法を拡張して，多次元空間においてエントロピーを最小化する境界面を見つけるようにすれば図 2 の赤線のカットポイントは抽出可能である．このカットポイントを得ることで，離散値においても領域 A と領域 B を正しく負例だと識別が可能となる．具体的には(1),(2),(3)式のエントロピーを最小化するカットポイントを全特徴量の中から探索して境界面を発見し，(4),(5),(6)式に従いカットが有効かを検証することで容易に拡張可能である．ただし，多数の特徴量で本拡張手法を実施すると，一部の有力な変数やそのカットポイントが支配的になり，離散化が一部の変数に偏る可能性がある．次元削減という観点では良いように見えるが，ルールベースの機械学習においては，経験則には無い新しい知識発見の可能性があるため，事前に特徴量を削りすぎるのは好ましくない．このようにルールベースの機械学習においては，特徴量が多い場合は特徴量の選択と本拡張手法を様々な特徴量パターンで実施することが必要となる．

3 提案手法

本提案手法のフローチャートを図 3 に示す．まず単変数の離散化を実施し，初期離散化データを作成する．そのデータに対して顕在パターンマイニングを実行する．得られたパターンの特徴量の組合せに応じて MDLP を拡張した複数変数離散化を実施する．パターンマイニングから離散化までの処理を設定回数繰り返す．

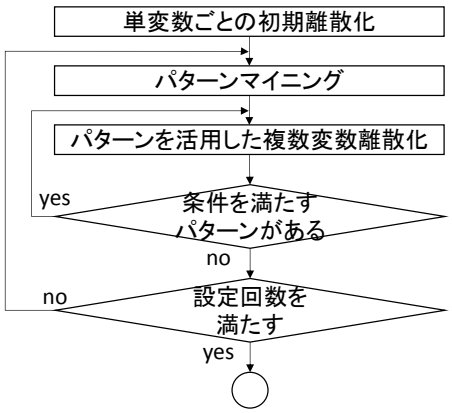

図 3 :提案手法のフローチャート

図 3 の個々の処理について個別に説明する.

3.1 単変数離散化

ここでは元のデータに対して単変数ごとに離散化を実施する．手法は等間隔離散化や 2.1 節で説明した MDLP による単変数離散化がある．今回応用するパターンマイニング手法に入力するための，初期離散化が必要となる．離散化後のカテゴリ作成には，値がそれぞれのカットポイントよりも下または上であることを表現した Ordinal Discretization の方式[6]を採用する．例えば年収 300 万円，700 万円をカットポイントとして抽出した場合，生成するカテゴリは「年収 300 万円未満」,「年収 300 万円以上」,「年収 700 万円未満」,「年収 700 万円以上」の 4 カテゴリとなる.

3.2 顕在パターンマイニング

入力データ項目から制約を満たす極小なパターンを網羅的に列挙する. 表 1 に wine データセットを使用して抽出したパターンの一部を示す．class 列はそのパターンが属するクラス，pattern 列は抽出されたパターン，support は class 列のラベルのカバー率，confidence はパターンに当てはまるデータのうちクラスが class 列と一致するデータの割合を示す.

3.3 パターンを活用した複数変数離散化

3.2 節で述べたように, パターンは制約を満たす極小なものとなる．このパターンで得られる変数の組合せを用いて複数変数離散化を実施する．表 1 の 1 行目であれば,「fixed acidity」と「density」,「pH」の 3 変数において 2.1 節で述べた複数変数に拡張した MDLP を実施する.

得られた全パターンの変数の組合せにおいて複数変数離散化によりカットポイントを取得し，OD の方式で離散化を実施する．また，パターンマイニングと複数変数離散化を繰り返し実行することで精度が向上するか検証する.

表 1 :wine データセットで抽出したパターンの一部

class	pattern	support	confidence
1	fixed acidity<10.3 ∧ density>=0.997 ∧ pH<3.058	0.0067	0.57
1	fixed acidity<10.3 ∧ residual sugar>=4.55 ∧ pH<3.058	0.0033	1.00
1	fixed acidity>=7.5	0.67	0.56
1	fixed acidity<7.5 ∧ total sulfur dioxide>=147.5	0.0033	1.00
1	fixed acidity<7.5 ∧ free sulfur dioxide>=18.75	0.12	0.55

4 実験結果

abalone, wine, banknote の 3 データセットにおいて，等間隔離散化，単変数の MDLP 離散化，本手法で比較をする.

4.1 精度比較

精度比較は表 2 の abalone, wine, banknote の 3 データセット[1]に対して行う．また, abalone データセットにおいては数値属性でない性別の変数を除外している．また, 2 値分類のためのラベルがないため，今回の実験では年齢が 10 以上かどうかをラベルとしてこれを予測する問題に置き換えている．wine データセットも同様に 2 値分類のためのラベルがないため, 評点が 6 以上かどうかをラベルとしてこれを予測する問題としている．これらのデータセットを使用して，顕在パターンマイニングで得られたパターンに対して Logistic Regression により線形分類モデルを生成する．また，他のアルゴリズムにも離散化が有効か検証するため，Random Forest(RF)でも精度測定を行う. 本実験では 20-fold での平均精度で比較を実施する.

表 2 :使用するデータセット

データセット	件数	列数
abalone	4177	8
wine	1599	12
banknote	1372	5

[1] https://archive.ics.uci.edu/ml/index.php

表 3 に初期離散化を等間隔離散化(ew:equal-width)と MDLP で実施した場合の結果を示す．init 列が初期離散化の手法を示す．初期離散化を実施後に，パターンマイニングと離散化を 5 回繰り返す．また，パターンによる分類と RF それぞれの精度も記載している．fscore(org)は初期離散化での F 値を，fscore(5iter)は本手法を 5 回繰り返した時の F 値を示している．どのデータセット，どの初期化手法においても本手法での精度が高いことがわかる．また，初期化手法は等間隔離散化で実施したほうが本手法の効果が大きいことがわかる．この理由として，MDLP の初期離散化ではすでに単変数ごとにラベルの相関を見ているため，複数変数でないと見えてこない関係性が欠落していることが考えられる．このことから，本手法における初期離散化はラベルの相関を考慮しない離散化手法の方が適していることが分かる．

表 3 :本手法を 5 回繰り返した時の精度

method	dataset	init	fscore(org)	fscore(5iter)
pattern	abalone	ew	0.7478	0.7828
pattern	abalone	mdlp	0.7715	0.7825
pattern	wine	ew	0.6944	0.7764
pattern	wine	mdlp	0.7538	0.7645
pattern	banknote	ew	0.9013	0.9982
pattern	banknote	mdlp	0.9268	0.9911
RF	abalone	ew	0.7485	0.7661
RF	abalone	mdlp	0.7534	0.7638
RF	wine	ew	0.6618	0.7635
RF	wine	mdlp	0.7427	0.7596
RF	banknote	ew	0.8977	0.944
RF	banknote	mdlp	0.916	0.9424

次に，繰り返し回数による精度の比較を表 4 に示す．初期化手法は，表 3 より equal-width に固定している．表より，繰り返し回数により精度が徐々に上がることは少ない．ただし，繰り返し回数が 1 回でも初期離散化より精度が上がるため，繰り返し回数を 1 回に設定しておけば安定的に精度向上することが見込める．

表 4 :繰り返し回数による精度の比較

	org	1iter	2iter	3iter	4iter	5iter
pattern abalone	0.7478	0.7703	0.7789	0.7834	0.7792	0.7828
pattern wine	0.6944	0.7701	0.7766	0.774	0.7757	0.7764
pattern banknote	0.9013	0.9934	0.9982	0.9982	0.9982	0.9982
RF abalone	0.7485	0.7545	0.7618	0.7689	0.7645	0.7661
RF wine	0.6618	0.7621	0.7542	0.7533	0.7588	0.7635
RF banknote	0.8977	0.9381	0.9412	0.944	0.944	0.944

5 まとめ

本研究では，高次のルールを扱う場合でも高精度を達成する離散化手法の実現を目的とし，既存の離散化手法と顕在パターンマイニングを組み合わせる手法を提案した．3 データセットにおいて，様々な条件の元提案手法の効果を検証し，精度向上できる離散化手法であることを示した．データセットによっては 10%に近い精度向上も見られた．

また，網羅的に列挙されたパターンを活用して離散化を実施することにより，単変数離散化では見逃してしまうルールの発見にも寄与できる．さらにモデルの良さを得られたルールに基づいて議論できるため AI-PoC の時間短縮につながると考える．

参考文献

[1] Biran Or, Cotton Courtenay : Explanation and justification in machine learning: A survey, IJCAI-17 workshop on explainable AI (XAI), Vol. 8, No. 1, (2017).

[2] Doshi-Velez, Finale, Been Kim: Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608, (2017).

[3] Ribeiro, Marco Tulio, Sameer Singh: Why should i trust you?: Explaining the predictions of any classifier, Proceedings of the 22nd ACM SIGKDD international conference on knopatternedge discovery and data mining, pp. 1135-1144, (2016)

[4] Milton García-Borroto, José Fco. Martínez-Trinidad, Jesús Ariel Carrasco-Ochoa: A survey of emerging patterns for supervised classification, Artif Intell Rev 42, pp. 705–721 (2014).

[5] 岩下，高木，鈴木，後藤，大堀，有村: 密なデータベースに対する動的な探索順序を用いた高速な顕在パターンマイニング手法，人工知能基本問題研究会 111, pp. 40-45, (2020)

[6] Yang Ying, Geoffrey I. Webb, Xindong Wu: Discretization methods, Springer, pp. 101-116, (2009)

[7] Garcia, Salvador, et al. "A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning." IEEE Transactions on Knopatternedge and Data Engineering, vol.25(4), pp. 734-750, (2013)

[8] U.M. Fayyad, K.B. Irani: Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning, Proc. 13th Int'l Joint Conf. Artificial Intelligence (IJCAI), pp. 1022-1029, (1993)

[9] E. Frank, I. H. Witten: Making better use of global discretization, Proceedings of the Sixteenth International Conference on Machine Learning, pp. 115-123, (1999)