カテゴリカル属性の自動判別方法の提案

Proposal of Automatic Discrimination Method of Categorical Attributes

鳴海　雄登† 　新美　礼彦†

† 公立はこだて未来大学システム情報科学部情報アーキテクチャ学科
〒 041–8655 北海道函館市亀田中野町
E-mail: †{b1016229,niimi}@fun.ac.jp

あらまし　近年，様々な分野においてデータマイニングに注目が集まっているが，データマイニングを行いたいと考える個人や団体が必ずしもデータマイニングの専門家であるとは限らない．その際，データを分析するためには外部へ委託，あるいはデータを分析できる人材を育成することになるが，どちらもコストが生じる．その際に生じるコストに対して，データを分析して得られる情報が見合っているかを事前に判断することは難しい．そこで本研究では，カテゴリカル属性の取り扱いに注目し，データセットのメタ特徴からカテゴリカルな属性とそうでない属性を判別する方法を提案する．提案手法を用い前処理を自動化することにより，データセットに有益な情報が含まれているかどうかを簡易的に判断できるシステムの構築が期待される．

キーワード　データマイニング，カテゴリカル属性，メタ特徴

1　はじめに

1.1　背　　景

近年，あらゆる分野においてデータマイニングに注目が集まっている．例えば，土木分野では橋梁のモニタリングにより膨大なデータが生成されるが，それらは “橋梁のある 1 点で変異の上ブレが観測された” などのデータに過ぎない．そのため石川らは複数のデータを総合的に見て，橋梁の一部に局所的に異常が生じている状態を抽出した [1]．この活用例のように，データ解析をする際には解析対象に関する知識とデータマイニングに関する知識の両方を持ち合わせている必要がある．

しかし，データマイニングを行いたい個人や団体が必ずしもデータマイニングの専門家であるとは限らない．その場合，データを分析するためにデータを解析する外部の機関や企業への委託，もしくはデータ分析を行うことのできる人材を育成する必要がある．そのどちらもデータを分析するための追加コストが，金銭や時間などで生じるが，そのコストを投じたことによって実際に有益な情報を抽出できるとは限らず，費用対効果をデータから判断することは難しい．そこで，データマイニングを行いたいユーザに対して，データマイニングに関する知識をシステム側で補うことができるシステムが有用であると考えられる．

1.2　簡易自動前処理システム

上述の背景を踏まえて，現在本研究では解析したいデータセットそのものと解析を行う方針から簡易的な前処理を自動的に行うことができるシステムの開発を目標にしている．本研究において前処理に着目した理由は，データマイニングのプロセスにおいて比較的初期段階であり，その後の他の作業へ大きな影響を与えやすいためである．

現段階で提案しているシステムは，カテゴリカル属性の処理を主軸とした簡易自動前処理システムである．データセットごとに各属性がカテゴリカル属性であるかを事前に指定しておき，特定のカテゴリの有無を表す 2 値の変数に変換する，ダミーコーディング (Dummy Coding) [2] をカテゴリカル属性に対して行うことによって，与えられたデータセットを自動的にデータマイニングを行うことのできる形に変形できるというものである．また，データ上値が空となっている欠損値は事例ごと削除 (List-Wise Deletion)，カテゴリカル属性として指定されなかった属性に含まれる数値以外のデータはエラー値として同様に事例ごと削除としている．

本研究におけるカテゴリカル属性の定義は，Stevens の尺度分類 [3] [4](表 1) のうち順序尺度 (Ordinal Scale) と名義尺度 (Nominal Scale) のいずれかに当てはまるものとした．すなわち，名義尺度，順序尺度は持つものの間隔尺度や比率尺度を持たないものを合わせてカテゴリカル属性と定義した．

本研究でカテゴリカル属性に着目したのは，本研究における定義に当てはまるカテゴリカル属性は数値ラベリングした際に，連続値として扱うことが適切とは言えないためである．いくつかのデータマイニングアルゴリズムにおいて，各属性を連続値として扱うことが求められるが，名義尺度のみをもつ属性は大小関係や間隔・差，比率の等値性が定義できないため，数値としてこれらを比較することが不可能である．また，順序尺度をもつ属性であっても，数値として大小関係の比較を行うことは可能であるが，それらの値間の間隔や差，比率について議論することは適当でない．そのためこれらの属性を連続値とは異なる取扱いを行えるように前処理を施すことによって，適切な扱いを行えるのではないかと考えられる．

実際に UCI Machine Learning Repository [5] より取得した

表 1　Stevens の尺度分類

尺度	基本的な経験的操作	数学的群構造	許容される統計量
名義	等値性の決定	置換群 $x' = f(x)$ $f(x)$ は任意の代入	事例数 最瀕値 偶発的な相関
順序	大小関係の決定	等方群 $x' = f(x)$ $f(x)$ は単調増加関数	中央値 パーセンタイル
間隔	間隔・差の 等値性の決定	一般線形群 $x' = ax + b$	平均値 標準偏差 順位相関 積率相関
比率	比率の 等値性の決定	相似群 $x' = ax$	変動係数

30 データセットに対し行った実験においては，前処理としてこのシステムを利用してカテゴリカル属性をダミー変数化した後，欠損値やエラー値を含む事例を取り除いたものを用いてデータマイニングを行うことはできた．

しかし，上述の通り，データマイニングの知識がない場合に，データマイニングアルゴリズム上でどの属性がカテゴリカルな性質を持ち，それは何故連続値とは異なる扱いをしなくてはならないのかという判断ができないという問題が生じ得る．そのため，各属性の性質からカテゴリカル属性を自動的に特定することができたならば，特定した属性をユーザへ提示し，この問題が解決できるのではないかと考えられる．

1.3 目　　的

本稿では，簡易自動前処理システムの構築を行う際に発生した問題に対する解決策のうちの一つである，各属性の性質からカテゴリカル属性を自動的に判断することを目的として，元のデータから抽出したいくつかのメタ特徴を用いて，特定の属性がカテゴリカル属性であるか否かを判別する学習器を構築する．

2 提案手法

各属性がカテゴリカル属性か否かが定かであるデータセットを複数用意し，各属性に関するいくつかのメタ特徴を抽出することにより学習データを作成する．

抽出したメタ特徴は，以下の 9 つである．

（1） 情報利得比
（2） 平均値
（3） 標準偏差
（4） 最小値
（5） 第 1 四分位数
（6） 中央値
（7） 第 3 四分位数
（8） 最大値
（9） ユニークな値の数

これらのメタ特徴を選択した理由を述べる．まず，情報利得比は決定木構築アルゴリズムである C4.5 [6] や CART [7] で用いられているためである．これらのアルゴリズムにおいて情報利得比は各属性により未分割事例を分割した際にどの程度クラス分類に寄与するかを計算する際に利用されるが，その際に情報利得比が高いほどその属性の分解能が高いと考えられる．先のアルゴリズムにおいて連続値に対して情報利得比がもっとも高くなるように閾値を設け分割した区間に対する情報利得比を用いるが，離散化せずに情報利得比を考えた際には，取りうる値の数と値 1 つあたりのクラス決定への寄与との比率がカテゴリカル属性に対して低くなるのではないかと考えられたためである．また，ユニークな値の数は事前にいくつかのデータセットに含まれるカテゴリカル属性の特徴を考察した際に多くのカテゴリカル属性がユニークな値が事例数に対して少ないと判断したため，それ以外の 7 つは統計量であるため，データセットの性質を調べるためにふさわしいのではないかと考えたためである．

次に，それらのデータセットの各属性がカテゴリカル属性か否かをクラスラベルとし，分類器を構築する．

表 2　LOO-CV の Confusion Matrix

-	Predicted_Negative	Predicted_Positive
Actual_Negative	429	18
Actual_Positive	20	477

3 実　　験

3.1 手　　順

分類器の評価には，学習データ全体に対する Leave-One-Out Cross-Varidation を行った結果と，学習に用いなかったいくつかのデータセットに対して予測を行った結果を用いる．

本実験では UCI Machine Learning Repository より取得した 37 データセットを用いた．クラスラベルは各データセットの説明文より作成した．なお，これらの用いたデータセットについては付録にて記載する．

本実験では，Scikit-Learn に含まれている Decision Tree Classifier(CART) を用いた．その際いくつかのパラメタを指定する必要があるが，本実験においてはすべてデフォルト値を用いた．

3.2 結　　果

分類器の Leave-One-Out Cross-Varidation の結果の Confusion Matrix を表 2 に示す．

その際の Accuracy，Precision，Recall を表 3 に示す．

また，本実験で構築した分類器における各メタ特徴の重要度は，最も重要なものがユニークな値の数，次点が情報利得比であった．

全訓練データで学習した決定木の 4 段目までを図 3.2 に示す．図中の X の添字 {0-8} は，それぞれ第 2 節において述べたメタ特徴の順番に対応している．この分類器の汎化性能を評価するために，いくつかのデータセットに対し分類を行った結果を表 4 に示す．

表 3　LOO-CV のスコア

指標	スコア
Accuracy	0.960
Precision	0.964
Recall	0.960

表 4　訓練データ外のスコア

データセット名	Accuracy	Recall	Precision	備考
Acute Inflammations	1.00	1.00	1.00	全正解
Arcene	0.983	-	-	負例のみ
Covertype	0.907	1.00	0.89	-
Gene Expression Cancer RNA-Seq	0.998	-	-	負例のみ
Poker Hand	1.00	1.00	1.00	全正解

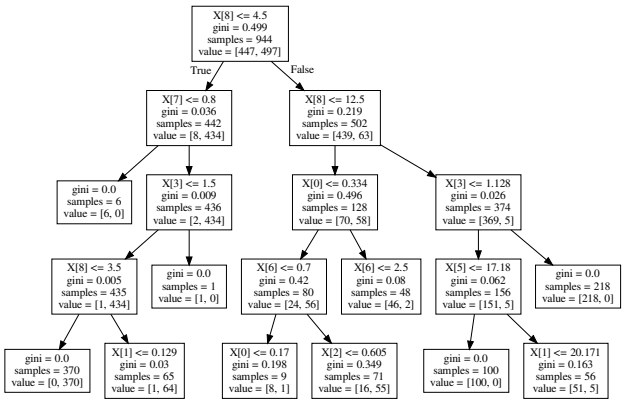

図 1　構築した決定木 (4 段目まで)

4　考　　察

本実験で用いたデータセットに対しては高精度で分類できることが確認できた．しかし，これは今回のカテゴリカル属性の定義を用いた上で，今回用いたデータセットにおいては分類できたということに過ぎない．

4.1　カテゴリカル属性の定義について

まず，今回のカテゴリカル属性の定義について再考する必要がある．第 1 節において，順序もしくは名義を示す属性を総称してカテゴリカル属性とするという一般的な考え方 [8] を示したが，その後の処理 (データマイニングにおけるデータの取捨選択やフォーマッティング，パタン発見，解釈等) においてカテゴリカル属性か否かを判定することによって自動的に適切なデータの取扱が行えるかは明確でない．そのため，データセットのメタ特徴から簡易的な前処理を自動的に行うという本研究の最終的な目標に向けて，特定の処理を行う必要のある属性を特定する場合には，更に詳細な分類が必要であると考えられる．

4.1.1　カテゴリカル属性について

本研究におけるカテゴリカル属性は，順序尺度と名義尺度のいずれかに当てはまるものとした．

本実験で使用した訓練データを用いて構築した決定木において，最重要視されているメタ特徴はユニークな値の数である．この決定木において第 1 段ではユニークな値の数に 4.5 という閾値を設けて事例をほぼ二分している (4.5 以下:442 事例，それ以外:502 事例)．そのうち，ユニークな値の数が 4 以下のもののほとんどが前述のカテゴリカル属性の定義に当てはまり (442 事例中 434 事例)，ユニークな値の数が 5 以上のものは大部分がこの定義に当てはまらないものであった (502 事例中 439 事例)．ここで，訓練データからいくつかの事例を取り出す．“Absenteeism at work” データセットに含まれる属性 “Reason for absence”，は欠勤の原因となった疾病を表す 21 のカテゴリと他 7 カテゴリの合計 28 カテゴリの内，各事例に当てはまるものが格納されている名義属性である．それに対し，“Balloons” データセットに含まれる属性 “size” は Large か Small かの 2 値格納されている順序属性である．

また，“Lymphography” データセットに含まれる属性 “no. of nodes in” はデータセットの説明においてはカテゴリカル属性とされているが，実数値を離散化したものである．しかし，最も大きな値が意味する離散値はある閾値以上の全ての数値の度数であるため，他の値の間隔と異なっている．そのため各値の間隔の等値性は満たしていないため，本研究におけるカテゴリカル属性の定義に当てはまり，順序属性であると考えられる．

これらの属性を一概にカテゴリカル属性として扱うということはその後の処理によっては適切でない場合もある．

4.1.2　非カテゴリカル属性について

本研究における非カテゴリカル属性は，上述のカテゴリカル属性の定義に当てはまらないものとしている．

本実験において汎化性能の評価に用いた，訓練データ外の “gene expression cancer RNA-Seq” データセットに含まれる属性 “gene_3527” のように 1 事例のみ非ゼロな実数値を持っており，それ以外の事例においては全てゼロを持つ属性や，同データセットの属性 “gene_5” のように全ての事例においてゼロである属性は，分析する際に連続値として扱うのが困難である場合もある．

また，“Poker Hand” データセットは説明文に従って作成したラベルに対しては全正解となった．このデータセットに含まれる属性 “Rank of card #1” や “Rank of card #5” などはトランプカードの数字 (Ace, 2, 3, ... , Queen, King) を表しており，説明文上では数値属性である．しかし，これらの属性はカテゴリカルな性質を持っているとも考えられる．トランプを用いたゲームの中にはカードの数字がそのカードの強さを表している場合もあるため，強さという数値が順序尺度や間隔尺度を持っていると考えられる．しかし，このデータセットにおいては，トランプカード 52 枚 (各スートにつき各数字のカード 1 枚ずつ，4 スート 13 組) のデッキの中から無作為に抽出した 5 枚のカードをそれぞれ “Suit of card #[1-5]” と “Rank of card #[1-5]” で表現しており，それらから構成されるハンド事例が 10 の役のうちのどれに対応しているかを目的属性としているため，このデータセットを用いた分類器を構築する際には各カード自体の強さは考慮されないと考えられる．そのためこれらの値は名義尺度のみを持っていると考えられ，本研究におけるカテゴリカル属性の定義に当てはまる．

これらの属性について本研究の定義を用いてカテゴリカル属

性ではないと判定することが，必ずしも適切であるとは言い難いと考えられる．

4.2 データセットについて

では，今回用いたデータセットに問題があるかを考える．一般に機械学習においては訓練データを増やすことによってカバーできるパタンが増え性能が向上すると考えられている．しかし，本実験の訓練データが対応するカテゴリカル属性のパタンを増加させても，この学習器の性能は向上しないと考えられる．

一般的な機械学習で扱うデータセットは多くが同一の母集団からサンプリングしたものである．そのため，データセットの事例数やクラスラベルをカバーできるパタンを増加させることによって，クラスラベルの定義が明確かつ典型的なパタンを発見する場合には，少数のデータを用いて学習する場合に比べて性能が向上する．

しかし，本実験で用いたデータセットは様々なデータソースからサンプリングされたデータセットから抽出したデータの集合であり，それらのデータソースは同一の母集団とは言えない．また，カテゴリカル属性という定義自体が，大まかな方針は共通認識としてあるものの，前述の通り分析対象やデータの性質によって同一なものになるとは限らない．そのため，より多くのパタンをカテゴリカル属性と判定するには，究極的には全てのパタンに対してカテゴリカル属性と判別してしまうということが考えられるが，それでは多様なデータセットに対して効率的かつ有用なデータの取扱いをすることは不可能である．

従って，前処理の自動化のために用いる特徴としてカテゴリカル属性という属性を用いるという考え方自体を再度検討し，前処理後のプロセスにおいて必要な特徴について再考することが必要であると考えられる．

5 おわりに

本稿では，名義尺度や順序尺度は持つものの間隔尺度や比例尺度を持たないものを合わせてカテゴリカル属性とする定義のもと，各属性の性質からカテゴリカル属性を自動的に判断するという目的のために，UCI Machine Learning Repository にて取得した 37 データセットより抽出した 9 つのメタ特徴を用いてカテゴリカル属性か否かを決定木によって分類した．

結果的に，各属性の性質からカテゴリカル属性を自動的に判断するという目的に対して，実験に用いたデータセット群に対しては高精度で分類を行えた．しかし，考察において述べたとおり，今回の実験ではデータセットの説明よりカテゴリカルか否かを示すクラスラベルを作成し分類を行ったが，データセットのメタ特徴から簡易的な前処理を自動的に行うという本研究の目標を達成するためには，前処理に用いる特徴について再考，あるいは細分化するべきではないかと考えられる．

今後は広範なデータセットに対して自動的に前処理を行うために，データマイニングにおいて特殊な処理が必要な属性はどのようなものを想定すべきであるかを再検討すると共に，どのような尺度を用いて属性区分を定義することによって適切な処理を行えるかを考察することを目標とする．

文　献

[1] 石川裕治, 宮崎早苗, “橋の異常を瞬時にキャッチ! - 橋梁モニタリングシステム BRIMOS の開発”, NTT 技術ジャーナル, Vol.21, No.9, pp.26-29, 電気通信協会, 2009.

[2] L.G. Grimm, P.R. Yarnold, “Reading and Understanding More Multivariate Statistics”, American Psychological Association, 2000.

[3] S.S. Stevens, “On the Theory of Scales of Measurement.”, Science, Vol.103, No.2684, pp.677-680, AAAS, 1946.

[4] 鷲尾隆, 元田浩, “尺度の理論”, 日本ファジィ学会誌, Vol.10, No.3, pp.401-413, 日本知能情報ファジィ学会, 1998.

[5] D. Dua, C. Graff, “UCI Machine Learning Repository”, http://arhive.ics.uci.edu/ml/, 2019.

[6] J.R. Quinlan, “C4.5: Programs for Machine Learning”, Elsevier, 2014.

[7] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, “Classification and Regression Trees”, CRC Press, 1984.

[8] 藤井良宜, “カテゴリカルデータ解析”, 共立出版, 2010.

付　録

1 実験に用いたデータセット

訓練データには以下のデータセットを用いた．

- Absenteeism at work
- Annealing
- Audiology (Standardized)
- Audit Data
- Balance Scale
- Balloons
- Blood Transfusion Service
- Breast Cancer
- Breast Cancer Coimbra
- Breast Cancer Wisconsin (Diagnostic)
- Car Evaluation
- Chess
- Congressional Voting Records
- Connect-4
- Cylinder Bands
- Divorce Predictors
- Glass Identification
- HCC Survival
- Habarman’s Survival
- Hayes-Roth
- Iris
- Lenses
- Lymphography
- MONK’s Problems
- Mammographic Mass
- Mushroom
- Nursery
- Primary Tumor
- SPECT
- SPECTF

- Shuttle Landing Control
- Soybean {Large, Small}
- Tic-Tac-Toe Endgame
- Ultrasonic flowmeter diagnostics {A, B, C, D}
- Vertebral Column {2C, 3C}
- Wine