卒業論文

データマイニングのための前処理アルゴリズム簡易自動選択システム

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
情報システムコース　1016229

鳴海 雄登

指導教員　　新美 礼彦

提出日　　2020 年 1 月 28 日

BA Thesis

A Simple Automatic Selection System for Pre-Processing Algorithms on Data Mining

by

Yuto NARUMI

School of Systems Information Science, Future University Hakodate
Information Systems Course, Department of Media Architecture
Supervisor:　　Ayahiko NIIMI

Submitted on January 28, 2020

Abstract–

Data mining has recently attracted research attention in all fields of study. Moreover, when analyzing data, it is necessary to know about the object and data mining. Therefore, in this study, we aimed to develop a tool to minimize the need for knowledge of data mining and facilitate effortless data mining. We focus on pre-processing that considerably influences the analysis results in data mining. In this study, we verified the type of processing performed on unprocessed data so that it can be automatically shaped and data mining can be performed on various data sets. Consequently, it is considered that categorical attributes and missing values should be carefully handled. Therefore, we should focus on the categorical attributes among them and consider them as categorical owing to the nature of data. We examined the method of automatic attribute discrimination. Finally, the categorical attributes could be automatically identified based on the definition used in this study, but it cannot not be stated that it could be solved from the viewpoint of properly grasping the meaning of the data. The policy to be investigated in the future s also discussed.

Keywords: Data Mining, Pre-Processing, Meta-Feature, Automatic Algorithm Selection

概 要：

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が，データを解析する際には解析する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．そこで本研究では，データマイニングに関する知識の必要性を最小限に抑え手軽にデータマイニングを行うことを補助するツールを開発することを目標とした． 本研究で着目したのはデータマイニングにおいて解析結果に大きな影響を与える前処理である．本稿では未処理のデータにどのような処理を行うことで，様々なデータセットに対してデータマイニングが行えるよう自動的に整形可能かを検証した．その結果，カテゴリカル属性と欠損値の取扱いについて注意すべきであると考えられたため，それらのうちカテゴリカル属性とは何かという部分に焦点を当て，データの性質からカテゴリカルであるとすべき属性の自動判別方法について検討を行った．最終的には，今回用いた定義の下ではカテゴリカル属性の自動判別を行えたが，データの意味を適切に捉えるという観点では解決できたとは言えないため，今後目指すべき方針について議論した．

キーワード： データマイニング, 前処理, メタ特徴, アルゴリズム自動選択

目次

第 1 章

序論

この章では，本研究の背景および目的を述べる．

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 点で変異の上振れが観測された” などの数値データに過ぎない．そのため石川らは複数のデータを総合的に分析し，橋梁の一部で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が必要となる．

しかし，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．その場合，データを分析するためにデータ分析を専門と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・時間共に生じるが，コストを投じたことで実際に有益な情報を抽出できるとは限らず，コストを投じる前に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，そのデータのメタ的特徴から分析の補助を行うシステムが有用であると考えられる．

1.2 データマイニングにおける前処理の重要性

データマイニングは大きく分けて，① データウェアハウシング，② 前処理，③ パタン発見（データマイニング），④ 解釈・評価の 4 つのステップによって成り立っている．データウェアハウシングは解析対象のデータを獲得・選択するプロセス，前処理はパタン発見を行うためにデータを整形するプロセス，パタン発見はデータからパタンを発見するプロセス，解釈・評価は発見したパタンを知識として活用するプロセスである．このデータマイニングのプロセスのうち，データの獲得，選択，前処理が最も重要と考えられる．データあっての

マイニングであり，マイニングの対象となるデータの質が悪ければ良い知識を発掘することはできない．

前処理プロセスの中には，ノイズ・外れ値・欠損値の処理，正規化・離散化，事例選択，属性選択，属性構築といった複数のプロセスが含まれている．これらはすべて，データセットの性質や解析方針によって最適な処理方法が異なる．前処理に不備があると，データマイニングによって有益な情報を得ることは困難となることは勿論のこと，データマイニングアルゴリズム自体を実行することができない場合もある．

1.3 目的

前述の通り，従来のデータマイニングではユーザがデータマイニングツールを使い試行錯誤しながらデータを分析することになるが，その中で前処理は分析結果に非常に影響を与える．また，データマイニングを行う際には用いるアルゴリズムに応じてデータを整形する必要が生じることもあるが，そのデータ整形もデータを獲得する際に行う場合もあれば，前処理において行う場合もある．しかし，データマイニングにおける試行錯誤では，背景において述べた通りデータマイニングに関する前提知識のほか，分析する対象に関しても知識を持ち合わせている必要がある．また，データマイニングのプロセスにおいて前処理は非常にコストが高い．そこで，データセットのメタ的特徴からそのデータセットに最適な前処理手法を提示し，それを適用することにより前処理の自動化が可能となるのではないかと考えられる．そのためこの研究では，生データを分析するために，ユーザの目的に合わせた前処理を自動的に行うという部分に焦点を当ててシステムを構築することを目標としている．

本研究は UCI Machine Learning Repository[2] より取得したデータセットを基に，生データにデータマイニングアルゴリズムを適用する際に最低限必要な前処理の自動化を行う．その際，はじめに大きな問題となったのは，欠損値・外れ値の扱いとカテゴリカル属性の扱いである．欠損値と外れ値は分析する際にノイズとなり得るが，何を以ってそれらの値を欠損値あるいは外れ値と判断するかという問題のほか，それらの値自体が意味を持つ場合もあるため単に削除するだけでは適当な扱いとは言えないという問題もある．またカテゴリカル属性は，Support Vector Machine(SVM)[3] や k-Nearest Neighbor(k-NN)[4, 5] などいくつかの統計的機械学習アルゴリズムにおいて数値として扱うことが求められるため，様々な方法を用いてアルゴリズム上で扱える形に変換する必要がある．

本稿ではこれらの問題のうち，カテゴリカル属性をデータマイニングで扱う際の自動前処理方法とデータの性質からカテゴリカル属性を同定する方法について提案することを目的とする．また，欠損値については一様にして欠損の含まれる事例を削除することにより対応した [6].

第 2 章

関連研究

この章では本研究を行うにあたって参考にした既存研究，あるいは関連する技術や手法について述べる．

2.1 データマイニングツールにおける前処理

この節では既存のデータマイニングツールにおける前処理機能について述べる．

近年，様々なデータマイニングに関わる手法が提案され，代表的なものはデータマイニングツールにおいて自動的に処理できるように組み込まれているケースも多い．そのため，データマイニングを行う際にあまり考えることなく使えてしまうため，データの意味や扱うべき形といったものが無視されているケースも多々見られる．

2.1.1 WEKA

WEKA(Waikato Environment for Knowledge Analysis)[7] は，Waikato 大学で開発された機械学習ソフトウェアである．WEKA は，ARFF 形式という属性の関係を包含するファイル，あるいは CSV 形式のファイルを読み取る．ARFF 形式では，“@RELATION <relation_name>” で関係の宣言，“@ATTRIBUTE <attribute_name> <datatype>” で属性名とそのデータ型の定義，“@DATA” はファイル内のデータセグメントの開始であるデータ宣言を示している．データ型は，数値 (Numeric)，名義 (Nominal)，文字列 (String)，日付 (Date) を使用でき，JDBC(Java DataBase Connectivity) を利用してデータベースからデータを取り込むことも可能である [8].

WEKA によってデータマイニングを行う際に，データセットに欠損値や外れ値 (ノイズ) を含むデータ，また一貫性のないデータについて前処理が必要となる．これらの前処理には “Filter” オプションが用意されており，ユーザは自分の行いたい処理のために必要な前処理を行うため，いくつかの “Filter” を組み合わせて前処理を行う．“Filter” の例として，欠損値に対しては “RemoveWithValues” を “matchMissingValues” に設定して欠損している

データを削除する，外れ値に対しては，“InterquartileRange” により四分位範囲を用いて外れ値と極値を含むインスタンスにタグ付けするなどが挙げられる．

2.1.2 RapidMiner

RapidMiner[9] は RapidMiner 社が開発・販売している商用データ分析ソフトウェアである．GUI を用いてデータの前処理，パタン発見，評価を行うことができる．データ分割や欠損値の補完などの前処理，決定木や単純ベイズ分類器などのパタン発見アルゴリズムの適用といった各機能がブロックで表現されており，それらをドラッグ&ドロップで配置し，アークで結合することによって分析プロセスを作成することができる．

2.1.3 Amazon SageMaker

Amazon SageMaker[10] は，Amazon 社が同社のサービスである AWS(Amazon Web Service) の一環として開発・提供している完全マネージド型の機械学習サービスである．様々な機能の複合体となっているが，その中で今回取り上げるのは Amazon SageMaker Autopilot である．

SageMaker Autopilot では自動的に未加工のデータを検証し，機能プロセッサを適用して，最適なアルゴリズムのセットを選出する．それは，自動的に作成されたいくつかの機械学習モデルからユーザが目的に合った最適なものを選択するというものである．その際に，モデルがどのように作成され，その中身がどうあるかなどが完全に見える化できるため，ユーザのモデルの精度およびレイテンシ等の要件を満たすモデルを選択するための材料が提供される．

2.1.4 既存ツールから見た本研究の立ち位置

本節で取り上げた既存ツールはいずれもデータマイニングを行う際にユーザの作業を補助するものである．1 つめに取り上げた WEKA はデータマイニングに関する研究においては典型的に利用されるツールであり，前述の RapidMiner や KNIME[11]，Pentaho[12] といったツールのデータマイニング処理ライブラリとしても広く用いられている．次に取り上げた RapidMiner，最後に取り上げた Amazon SageMaker はいずれも商用システムである．

RapidMiner や Amazon SageMaker 等の商用データマイニングシステムにおいては，いずれも機械学習の経験のないユーザにも使用が可能であるとされているが，それは本来行うべき作業を抽象化しているのであって，データマイニングに関する知識がない状態で運用することは難しいと考えられる．

本研究のユースケースであるデータマイニングの知識がないユーザというのは，これらのデータマイニングツールが指す初学者とは異なる．本研究が目指すシステムは，よりデータマイニングに関する知識がない状態で利用でき，ユーザの本来の目的に対して大まかに利用

可能であるかを判断するために利用し，システムによりデータの価値を最低限見出すことができればより専門的な解析を行うためにコストを投じるというものである．そのためこのシステムは，用いることでユーザがデータマイニングにより達成したい目的を直接達成できるというものではなく，用意したデータ，あるいは検討しているデータがその目的を達成するための要件を満たしているかを確認し，その後に生じるコストについて費用対効果を検討するためのものである．

2.2 データセットのメタ特徴を用いたアルゴリズムの自動選択

この節ではデータセットのメタ特徴を用いたアルゴリズムの自動選択に関する関連研究について述べる．

2.2.1 データセットの特徴選択の自動化

Filchenkov らはデータセットのメタ特徴を用いて特徴選択を行った [13]．データセットの小さな部分サンプルによる処理アルゴリズムのパラメタの最良推定に基づくメタ特徴選択手法により，16 の特徴選択アルゴリズムに対し，彼らが提示した 44 のメタ特徴の中から，13 のメタ特徴を用いた Wang らの研究結果 [14] と比較して，18 のメタ特徴により 5 つのデータマイニングアルゴリズムに対して F 値を向上させた．

2.2.2 データマイニングアルゴリズム選択の自動化

南保らはデータセットのメタ特徴を用いてデータマイニングアルゴリズムの自動選択を行った [15]．54 種類のメタ特徴に対し学習データの部分特徴集合をサンプリングし，選択された特徴を用いて識別機を構築し分類精度を求めるという操作を，全ての部分集合に対して繰り返し，分類精度が最大となるときの特徴の集合を選択するという手法により，26 のメタ特徴を用いた Nakamura らの研究結果 [16] と比較して，5 つのメタ特徴により F 値を向上させた．

2.2.3 関連研究と本研究の関係

上述の関連研究はいずれもデータセットのメタ特徴を用いてデータマイニングにおける各プロセスにおいて作業を自動化するという部分で本研究と類似している．しかし，Filchenkov らの研究では特徴選択を，南保らの研究ではデータマイニングアルゴリズムの選択を自動化している点で本研究と異なる．

本研究ではこれらの関連研究を参考に抽出するメタ特徴の方針を決めようと考えたが，本研究の自動化対象は前処理であるため，対象が異なることによりメタ特徴の流用は不可能であった．

2.3 カテゴリカルデータの取扱い

この節ではカテゴリカルな性質を持ったデータの取扱いに関する関連研究について述べる．

2.3.1 測定尺度の理論

Stevens は測定の尺度を特定のクラスに分類されるという内容について報告した [17].

測定とは，広義にはルールに従ってオブジェクトまたはイベントに数字を割り当てることとして定義されている．その際に異なるルールの下で数字を割り当てることができるが，異なる種類の尺度と異なる種類の測定によって，数字の割当に関する様々な規則，結果の尺度の数学的特性の違い，尺度の各タイプで行われた測定に適用可能な統計演算が明確とならないといった問題が生じていた．そこで測定された値を，測定プロセスで呼び出される経験的操作と尺度の正式な数学的プロパティの両方によって決定されたクラスに分類することで，データに対して安全に適用できる統計的操作を，順序付けられた尺度タイプにより決定するというアプローチで問題解決を試みた．

2.3.2 数量化と主成分分析

足立は，多変量解析においてカテゴリカルデータを扱う際に行う数量化の諸体系の中でも，Gifi システム [18] と総称される等質性分析およびその関連手法について報告した [19].

等質性分析とは，複数の名義尺度に対して関係性を示すために用いられる手法である．等質性分析では各事例のスコアとその事例が該当するカテゴリのスコアは近く，等質的な値を取るべきであるというスコア間の等質性の仮定を基にしている．ここで，スコアというのは，数量化における事例とカテゴリに付与される数量的得点のことである．数量化は，各カテゴリカル変数につき事例ごとに取るカテゴリのデータ行列から，何らかの基準によって最適なスコアの行列を求めることである．等質性分析では，カテゴリカル変数 j について事例 i が反応したカテゴリ h_{ij} の第 s 次元の得点 $\hat{y}_{jh_{ij}s}$ の $i = 1, \cdots, n$ についての分散 η^2_{js} を，変数 j の次元 s における判別測度 (Discrimination Measure) と呼ぶ．この判別測度は次元 s の得点が変数 j のカテゴリ間の識別にどの程度寄与しているかを表し，変数 j と次元 s の関連度の指標となる．また，等質性分析によって得られた解 (前述の最適なスコアの行列) の各行ベクトルを点とみなして，第 s 次元の座標値としてプロットすると，個体およびカテゴリの空間布置が得られる．これらを用いて，類似した性質を持つカテゴリを推論することができる．

等質性分析のうちの一つである非線形主成分分析は，m 変数の中のいくつかについて，カテゴリ得点行列の階数を 1 と制約したものである．この制約により，変数内のカテゴリが原点を通る直線つまり 1 次元の尺度上に位置づけられ，この 1 次元の尺度上の長さを 1 次元スコア (Single Quantification) と呼ぶ．この制約は，設問 (変数) に対し段階的な解答 (カテゴリ) がされる場合に各カテゴリが一次元上にあると仮定できる場合に妥当と言える．そ

のため非線形主成分分析を用いる際には，制約を課さない変数，一次元性だけを仮定する変数，順序性を仮定する変数，等間隔性を仮定する変数として解析対象とする．これらはそれぞれ，多次元名義 (Multiple Nominal)，一次元名義 (Single Nominal)，順序 (Ordinal)，数値 (Numerical) 変数と呼ぶ．非線形主成分分析においても，等質性分析において述べた解釈方法を同様に用いることができる．また多次元名義変数において相関係数が次元と変数の相関の二乗を表すため，各変数の各次元に対する相関を推論できる．

2.3.3 関連研究と本研究の関係

本研究においてカテゴリカル属性を定義する際に，Stevens の論文を参考にした．この尺度分類は一般的に用いられており [20]，本研究においてもそれを踏襲して定義を行った．しかし，詳細は実験考察において述べることとするが，本研究の目標であるデータマイニングにおける前処理の自動化という観点においては，後述のカテゴリカル属性の定義を用いるだけでは，データセットの各値に対する適切な取扱いとは言い難い側面もあるため，今後の検討事項である．

足立の論文で紹介されている Gifi システムについては，各カテゴリカル変数の相関や，カテゴリカル変数が成す次元についての相関を考察するために役立つ技術ではあるが，本研究の目的であるデータマイニングにおける前処理の自動化という観点においては，目的が異なっているため参考としていない．しかし，非線形主成分分析における変数の取り扱いで，名義尺度を多次元名義尺度と一次元名義尺度という形で分離しているのは，前述の通りデータセットの各値に対する適切な取扱いについて議論する際に有用なのではないかと考えられる．

第 3 章

提案手法

この章では提案手法について述べる．

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．第 1 章の目的において述べた通り，前処理の自動化を行う際の問題点のうちの一つを解決するアプローチとして提案するのが “特定の性質を持つデータセットに対する自動前処理” である．

しかし，背景において述べた通り，本研究ではデータマイニングの知識がない場合をユースケースとして想定している．ユーザにデータマイニングの知識がない場合，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．この問題解決を行うためのアプローチとして提案するのが “データセットのメタ特徴を用いたカテゴリカル属性の判別” である．

3.1 特定の性質を持つデータセットに対する自動前処理

前処理を行っていない生データに対し，データマイニングアルゴリズムを実行するに当たって必要な前処理を試行錯誤的に適用した結果から，同様の前処理が必要となるデータセットを推定し，前処理を行うこととする．本稿で提案する方法は，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，カテゴリカル属性を，特定のカテゴリの有無を表す 2 値の変数に変換する，ダミー変数化 [21] を行うことによって，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．ダミー変数の利用例として，行動科学や社会科学等の分野において多く用いられている統計手法の一つである重回帰分析と相関分析 (Multiple Regression and Correlation, MRC) では，連続尺度や比率尺度で測定した量的変数の他に，多くのカテゴリカルな変数に対応させるためにカテゴリカル変数の変換が必要となる．その際に用いられる手法としてダ

ミーコーディングがある．例えば，Willshire らの研究 [22] において，性別はカテゴリカルな変数であり，2 つのカテゴリ (男性，女性) に対し男性を 1，女性を 2 とした．なお，ここで女性に対して大きな値を与えているがそれは量的な解釈ができるものではないことに注意が必要である．このように，ダミー変数はカテゴリカルな属性を持つものに対し，アルゴリズムの性質上数値として扱う場合に用いられる．

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は不適当である．例えば，日本において出身地を述べる際に，都道府県は 47 カテゴリを持つカテゴリカルデータと言える．それらに数値ラベリングを行った際に，北海道を 1 としてそこから順に数字を割り振り，沖縄が 47 という場合が考えられる．その場合に，2 である青森は北海道より優れているというような順序性や，平均値を求めるなどの操作は認められない．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．例えば，多くの商品レビュー等で 5 段階評価を行うことがあるが，その際の 4 という評価は 2 という評価に比べて 2 倍である，また 3 と 4 の差は，4 と 5 の差と等しいなどの議論は適当とは言えない．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．例えばダミー変数を用いて，前述の都道府県の例においては，北海道から沖縄までの 47 変数においてそれぞれ該当する変数に 1，それ以外を 0 としてデータを生成することによって，前述の問題を回避できると考えられる．

また，本研究では性能評価としてデータマイニングアルゴリズム実行後の識別精度向上ではなく，生データをデータマイニングアルゴリズムを実行可能にするための前処理を自動化することを当面の目標としている．

3.1.1 本研究におけるカテゴリカル属性の定義

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 [17, 20](表 3.1) 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例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．

3.2 データセットのメタ特徴を用いたカテゴリカル属性の判別

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．

表 3.1 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

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[23] や CART[24] 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 (あるいは CART においては GINI Index) 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 つは基本統計量であるため，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られるためである．基本統計量は，データの性質を調べるためにまず調べられる特徴量である．そのため，基本統計量がカテゴリカル属性の判別に有効であるかを確認するため，メタ特徴として取り上げた．

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．

データセットの性質からカテゴリカル属性であるかを判別することによって，上述の通り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を回避することが期待される．それにより，カテゴリカルな性質を持つデータに対し，単にそれらを数値として扱うことを防ぐことができると考えられる．

第 4 章

実験と評価および考察

この章では本研究で行った実験と評価および考察について述べる．

以下の実験は，それぞれが提案手法に対応している．実験 1，2 は提案 “特定の性質を持つデータセットに対する自動前処理”，実験 3 は提案 “データセットのメタ特徴を用いたカテゴリカル属性の判別” と対応している．

実験 1 では自動前処理システムを構築する必要性を検証するために実験を行い，実験 2 では実験 1 においてデータマイニングアルゴリズムを実施する際に必要な最低限の前処理を自動化した際に発生した問題に対する解決方法を実験によりデータを用いて検証した．実験 3 では，実験 2 において手動で指定していたカテゴリカル属性を自動判別する方法について実験で検証した．

4.1　実験 1 － 前処理の必要性に関する検証

実験 1 は，本研究の必要性を確認するために実施した．データマイニングでは一般的に前処理を行うが，実際に用いられているデータに対しデータマイニングアルゴリズムを実行する際にどのような前処理を行うかは分析対象のデータ毎に異なる．それらの前処理の中には，データマイニングアルゴリズムを実行する際に識別精度 (正解率や適合率，再現率など) を向上させるために行っているものも含まれている．そこで，本研究の目標である，生データにデータマイニングアルゴリズムを適用する際に最低限必要な前処理の自動化を行うために，生データに対しどのような前処理を行うことでデータマイニングが行えるかを検証する．

4.1.1　実験目的

生データに対して前処理の自動化を考えた際に，実際にどのような前処理が必要であるかを検討する必要がある．そのため本実験では，未処理の生データに対し分類データマイニングを行う際に最低限必要な前処理を特定する．

表 4.1 エラーが発生した例のメタ特徴

データセット名	含まれる属性の種類	事例数	属性数
Audit Data	Real	777	18
HCC Survival	Integer,Real	165	49
Mammographic Mass	Integer	961	6
Annealing	Categorical,Integer,Real	798	38
Cylinder Bands	Categorical,Integer,Real	512	39
CongressionalVotingRecords	Categorical	435	16
Primary Tumor	Categorical	339	17
Tic-Tac-Toe Endgame	Categorical	958	9

4.1.2 実験手順

本実験では，UCI Machine Learning Repository より取得した 30 データセットに対して次の操作を行った．なお，用いたデータセットについては付録にて記載する．

配布されているデータセットに対し，前処理を行わずにデータマイニングアルゴリズムを適用した．その後，処理できなかったデータセットに対して原因を追求し，それを解決する前処理を実行した．その際に同一の前処理によってデータマイニングアルゴリズムを適用可能になったものについて共通の特徴を調べた．

また本実験において，このリポジトリにてデータセットの絞り込みを行う際に使用される特徴をデータセットを分類する際に有効なメタ特徴であると仮定した．

実験において用いたデータマイニングアルゴリズムは，“The Top Ten Algorithms in Data Mining”[25] を参考に，CART，Support Vector Machine，Naive Bayse，3-Nearest Neighbor である．また，分割数 5 の交叉検証で正答率の平均値を評価した．なお実験には Python を用い，データマイニングアルゴリズムの実行には Scikit-Learn，データの読み込み・取扱は Pandas を用いて行った．

4.1.3 実験結果

30 データセット中 22 データセットについて前処理を行わずにデータマイニングアルゴリズムが実行でき，その内 1 データセットで正答率が著しく低い結果となった．

エラーの発生した例について，メタ特徴を比較した結果を表 4.1 に示す．

これらの例のうち，Attribute Type に着目してエラーの発生要因を検証した．まず，8 例中 5 例がカテゴリカル属性を含んでいるため，カテゴリカル属性を含み成功した例と比較を行う．本実験に用いたデータの内，Attribute Type がカテゴリカル属性のみかつ事例数が 100 から 1000 であり特徴数が 10 から 100 の例の実行結果を表 4.2 に示す．

ここに，NB は Naive Bayes，(G)，(B)，(M) はそれぞれ Gaussian 分布，Bernoulli 分布，

表 4.2 該当する例

データセット名	CART	SVC	NB(G)	NB(B)	NB(M)	3-NN(u)	3-NN(d)
CongressionalVotingRecords	-1	-1	-1	-1	-1	-1	-1
Lymphography	74.8	79	68.6	54.7	78.3	73.7	73.7
PrimaryTumor	-1	-1	-1	-1	-1	-1	-1

Multinomial 分布を表している．3-NN は 3-最近傍法を表しており，(u)，(d) はそれぞれ重みを均等につけたもの，距離に応じて変化させたものを表している．表中の-1 はエラーを表していて，正の実数値は平均スコア (%) である．

この 3 例について，予め想定していたメタ特徴のみでは切り分けることができないため，各データセット内の事例の比較を行った．その結果，データマイニングアルゴリズムの実行に成功した “Lymphography” データセットにおいてカテゴリカル属性に整数が割り当てられていたが，他 2 例においては文字列が割り当てられていた．エラーが発生した例のうち，Attribute Type にカテゴリカル属性を含む他 2 事例のほか，“Audit Data”，“HCC Survival”，“Mammographic Mass” に関しても UCI データリポジトリの記述においては Attribute Type にカテゴリカル属性を含んでいないとされていたが文字列も含んでいたため同条件を満たしていた．

また，これら 8 例について，文字型ラベルを数値ラベルに置換した後においても “Audit Data”，“Cylinder Bands” についてエラーが発生した．この 2 例についても予め想定していたメタ特徴のみでは他例と切り分けることができないためデータセット内の事例の比較を行ったところ，2 例ともに欠損値を含んでいるという点で他データセットとの差異があった．

4.1.4 考察

本実験の目標である，未処理の生データに対し分類データマイニングを行う際に最低限必要な前処理の特定は達成された．

本研究の第 1 目標を達成するために，エラーが発生しデータマイニングアルゴリズムが実行できなかったデータセットに対して前処理を行う必要がある．今回用いたアルゴリズムの仕様上，文字型ラベルや欠損値を含んでいる場合には，処理が行えないケースがある．そのため，これらを解決するために以下の 2 つの対策を講じた．

1 つ目の対策は，各特徴に対しユニークな値を取得し，文字型ラベルが含まれている場合に数値への置き換えを行うというものだ．2 つ目の対策は，各事例に対し欠損値を含んでいるものを探し，欠損値が含まれている事例を削除するというものである．これらの対策により，エラーが発生していたデータセットに対してもデータマイニングアルゴリズムを実行することが可能となった．

また，UCI データリポジトリのデータを利用する際に，記載されている情報と実際のデー

タセットの内容との間に差異が見られたので，以後実験を行う際には対象データと表記に関する検証を行う必要があると考えられる．

4.2 実験 2 ー カテゴリカル属性指定による自動前処理の有効性の検証

実験 2 は，実験 1 における対策について後述の問題が発生していたため，それを解決するために実施した．統計や多変量解析等においてカテゴリカルデータの取扱いは様々な手法が提案されている [26]．そのためカテゴリカルデータを取り扱ういくつかの手法が存在するが，本実験ではすべてのカテゴリカル属性に対しダミー変数化を行うことで適切な取扱いができるのではないかという提案を検証する．

4.2.1 実験 1 における問題点

実験 1 において講じた対策にはいくつかの問題が存在している．

まず，データセット上文字等で表されているもののほか，数値データとして値が入力されているものも存在するカテゴリカル属性の取扱だ．文字等のラベルで表されているデータに関しては登場順に数値に置き換える処理を行っており，データセットにおける値の表現としては数値データとして入力されているものと同等である．しかし，これらのデータについて数値として取り扱う場合には問題が生じ得る．

例として，カテゴリカル属性のみからなるデータセット “Lymphography” に対し実験 1 において用いたデータマイニングアルゴリズムを用いる場合を考える．“Lymphography” データセットは 18 のカテゴリカル属性 (表 4.3) から多値分類を行う．

カテゴリカル属性が数値で表されていることを一切考慮せず，Scikit-learn にて分割数 5 の交叉検定を行った Accuracy を表 4.4 に再掲する．

なお，CART は MAX DEPTH が 5，k-NN の k を 3，全てのデータマイニングアルゴリズムのそれ以外のパラメタはデフォルト値である．

“Lymphography” データセットでは，2-8, 16, 17 は 2 値属性であるが，それ以外は多値属性である．多値属性の中で，9, 10, 18 はそれぞれリンパ節の縮小と拡大，離散化した節の数を数値で表現しているので，少なくとも大小関係は保証されていると考える．しかし，9, 10, 18 以外の多値属性はそれぞれに大小関係があるとは考えにくい．

各データマイニングアルゴリズム上での扱いを考える．まず CART(Classification And Regression Tree) において決定木を構築する際には，すべての属性を連続値として，ジニ係数や情報エントロピーに基づいて効率が最大になる閾値を設ける．そのため，大小関係が明確でない名義属性においては適切な扱いとは言い難い．SVM(Support Vector Machine) と k-NN(k-最近傍法) では距離計算を行うため，間隔や差の等値性が明確でない順序属性や名義属性においては適切な扱いとは言い難い．Naive Bayes では連続値を扱う際には特定の確

表 4.3 “Lymphography” の説明属性

カラム No.	属性名	取りうる値	値に対する意味
01	lymphatics	1, 2, 3, 4	normal, arched, deformed, displaced
02	block of affere	1, 2	no, yes
03	bl. of lymph. c	1, 2	no, yes
04	bl. of lymph. s	1, 2	no, yes
05	by pass	1, 2	no, yes
06	extravasates	1, 2	no, yes
07	regeneration of	1, 2	no, yes
08	early uptake in	1, 2	no, yes
09	lym. nodes dimin	1, 2, 3	1, 2, 3
10	lym. nodes enlar	1, 2, 3, 4	1, 2, 3, 4
11	changes in lym.	1, 2, 3	bean, oval, round
12	defect in node	1, 2, 3, 4	no, lacunar, lac. marginal, lac. central
13	changes in node	1, 2, 3, 4	no, lacunar, lac. marginal, lac. central
14	changes in stru	1, 2, 3, 4, 5, 6, 7, 8	no, grainy, drop-like, coarse, diluted, reticular, stripped, faint
15	special forms	1, 2, 3	no, chalices, vesicles
16	dislocation of	1, 2	no, yes
17	exclusion of no	1, 2	no, yes
18	no. of nodes in	1, 2, 3, 4, 5, 6, 7, 8	0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, >=70

表 4.4 未考慮の正解率

CART	SVC	NB(G)	NB(B)	NB(M)	3-NN(u)	3-NN(d)
74.8	79.0	68.6	54.7	78.3	73.7	73.7

率分布を仮定して分類器を構築するため，カテゴリカルな多値属性は連続値として扱われているため適切な扱いとは言い難い．

今回想定した “Lymphography” データセットでは 18 個のカテゴリカル属性のうち 9 個の属性が多値属性であった．このデータセットでは元々カテゴリカル属性が数値によって表現されていたため，実験 1 において目的としていたデータマイニングアルゴリズム実行時のエラー解消という点において何も対策していなかった．また，前述のとおりカテゴリカル属性を数値化するという対策を講じたデータセットにおいても，同様の問題が生じていると考えられる．よって，実験 1 において施した前処理は，各属性のアルゴリズム上での取扱いを考えた場合において適切でないと考えられる．

次に，欠損値の取扱である．現状では欠損値は，何かしらの値として表現されている場合は文字等でラベル付されているものは数値化し，数値として入力されているものはそのまま用い，データセット上で欠損しているものは事例ごと削除している．しかし，欠損値を含む事例の削除によって小規模なデータであれば事例数が更に減ってしまうほか，数値として欠損値が表現されている場合には不適切なデータの取扱となり，更に欠損値は欠損値自体に意

味がある場合も考えられる．数値として表されている場合の問題について，例えば慣例的に用いられることの多い欠損値を-1 という数値により表現するという場合が該当する．その場合，上述の CART によって決定木を構築する際に欠損値を表す-1 も実数として扱われてしまうため，事例全体のうち欠損値が多く含まれている場合には計算結果に影響を与えてしまう．欠損値自体に意味がある場合というのは，例えば健康診断データを分析する際に必要の無い検査項目については検査を行わないため欠損値となるが，それは機材の故障や人為的ミスによってデータが欠損しているわけではなく，欠損値には検査の必要がなかったという意味が付与されることとなる．

4.2.2 実験目的

実験 1 の問題点を受けて，本実験ではカテゴリカル属性の取扱方法について検証する．本実験では予め各属性についてカテゴリカル属性であることを示すメタデータを作成しておき，それを基にカテゴリカル属性の自動フォーマットをした場合にカテゴリカル属性を適切に扱うことができるか確認することを目的とする．

本実験において作成したメタデータは以下の内容である．

- データセット名 (識別用)
- データセットのディスク上の位置 (絶対パス)
- データセットのファイル形式 (今回は CSV のみ)
- データセットファイルの CSV の区切り文字
- データセットファイルがヘッダを含むか否か
- データセットファイルがインデックス列を含むか否か
- データセットファイルのインデックス列の列番号
- データセットファイルのターゲット列の列番号
- データセットがカテゴリカル属性を含むか否か
- データセットファイルのカテゴリカル属性の列番号 (リスト)

メタデータの一例として，“Lymphography” データセットの例を付録に記載する．

4.2.3 実験手順

本実験では，実験 1 で用いたものと同様のデータセットを用いて，カテゴリカル属性を全てダミー変数化することによって分類データマイニングを行えるかを検証した．ダミー変数化 (Dummy Coding) は Willshire ら (1991) の研究 [22] 等で用いられているが，多変量解析においてカテゴリカルな変数，あるいは名義尺度水準の変数に対応させるために用いられる，特定のカテゴリの有無を表す 2 値の変数に変換するというものである [21].

UCI Machine Learning Repository におけるデータセットの説明文から 3 章におけるカ

テゴリカル属性の定義に当てはまる属性をカテゴリカル属性とし，それらの属性がカテゴリカル属性であると示すメタデータを作成する．その後，そのデータを基にカテゴリカル属性をダミー変数化し，非カテゴリカル属性に含まれる文字型ラベルをエラー値として，エラー値と欠損値を含む事例を全て削除する．これらの処理が終了したデータセットに対して実験 1 と同様のデータマイニングアルゴリズムを適用した際に，アルゴリズム上でカテゴリカル属性が適切に扱われているかを確認する．なお実験には Python を用い，データマイニングアルゴリズムの実行には Scikit-Learn，データの読み込み・取扱・ダミー変数化は Pandas を用いて行った．

4.2.4 実験結果

30 データセット全てにおいてデータマイニングアルゴリズムが適用できた．

4.2.5 考察

本実験の目的である，カテゴリカル属性の自動フォーマットをした場合にカテゴリカル属性を適切に扱うことができるかの確認は以下の作業により確認した．

ここでは，“実験 1 における問題点” にて用いた “Lymphography” データセットを用いてカテゴリカル属性の扱われ方について示す．

“lymphatics” 属性において元々の値は 1,2,3,4 であり，CART により決定木を構築する際に，これらを実数として用いた際にターゲット属性に対するジニ係数が最大となるように分割する際に閾値が 1.5 となる．この後も繰り返し分割することで，最終的には分割しきれない葉のジニ係数が 0.487,0.500,0.552 となるが，途中で用いられる閾値は 3.5,2.5 と値自体にはあまり意味が見られない．この属性をダミー変数化した際，同様にジニ係数が最大となるように分割する際に閾値は 1 に対して 0.5 となるが，これは 1 に該当するものと該当しないものを分割しているため値自体に意味があると考えられる．その後の分割も 4 に対して 0.5 と 2 に対して 0.5 となり，これらも同様に 4 に該当するか否か，2 に該当するか否か，という意味になる．SVM，k-NN の距離計算に関しては，1,2,3,4 のまま運用する際には本来の意味である normal,arched,deformed,displaced がそれぞれ隣接しているものが近いという計算になるため適切な扱いとは考えられないが，これらをダミー変数化した際にはそれぞれの値について該当するか否かが 0,1 で入っているため同値であれば距離 0，そうでなければ距離 1 となる．

このように，実験 1 と比較してカテゴリカル属性の扱いは適切なものとなっていることが確認できたが，上述の通りエラー値と欠損値を含む事例を削除していることによって，事例数が激減したものもあった．“HCC Survival” データセットにおいて，事例削除前の事例数は 165 であったが当該事例を削除したことによって事例数が 8 に減少している．このようにエラー値や欠損値の多いデータセットに対して自動前処理を行う際にはエラー値や欠損値の

取扱いについて再考する必要がある．また，今回はカテゴリカル属性か否かを示すメタデータをデータセットの説明から手動で作成したが，本研究のユースケースであるデータマイニングの知識がない場合を考えた際に，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つのかを判断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，ユーザへの提案ができ，この問題を解決できるのではないかと考えられる．

4.3 実験 3 － カテゴリカル属性の自動判別法の検証

実験 3 は，カテゴリカル属性を自動判別する方法について考察するために行った．実験 2 において各属性がカテゴリカル属性か否かというラベル付けを手動で行っていた．本研究の目標であるシステムのユースケースにおいてデータマイニングの知識を有していないユーザも考慮する必要があるため，データマイニング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．その際にユーザに対しカテゴリカルな性質を持つ属性を提示するために本実験を行う．

本実験は DEIM2020 に “カテゴリカル属性の自動判別方法の提案” の題で投稿し，第 12 回データ工学と情報マネジメントに関するフォーラムにおいて発表予定である．

4.3.1 実験目的

本実験では，実験 2 で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．

4.3.2 実験手順

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した．抽出したメタ特徴は第 3 章 2 節の通りである．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．

分類器の評価には, 学習データ全体に対する Leave-One-Out Cross-Varidation(LOO-CV) 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，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

表 4.5 LOO-CV の Confusion Matrix

-	Predicted_Negative	Predicted_Positive
Actual_Negative	429	18
Actual_Positive	20	477

用いた.

なお実験には Python を用い，データマイニングアルゴリズムの実行には Scikit-Learn，データの読み込み・取扱・統計量の抽出・ユニークな値の集計は Pandas，情報利得比の計算は Numpy と Pandas を用いて行った.

4.3.3 実験結果

分類器の Leave-One-Out Cross-Varidation の結果の Confusion Matrix を表 4.5 に示す.

その際の Accuracy，Precision，Recall を表 4.6 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 4.1 に示す．図中の X の添字 {0-8} は，それぞれ第 3 章 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 4.7 に示す.

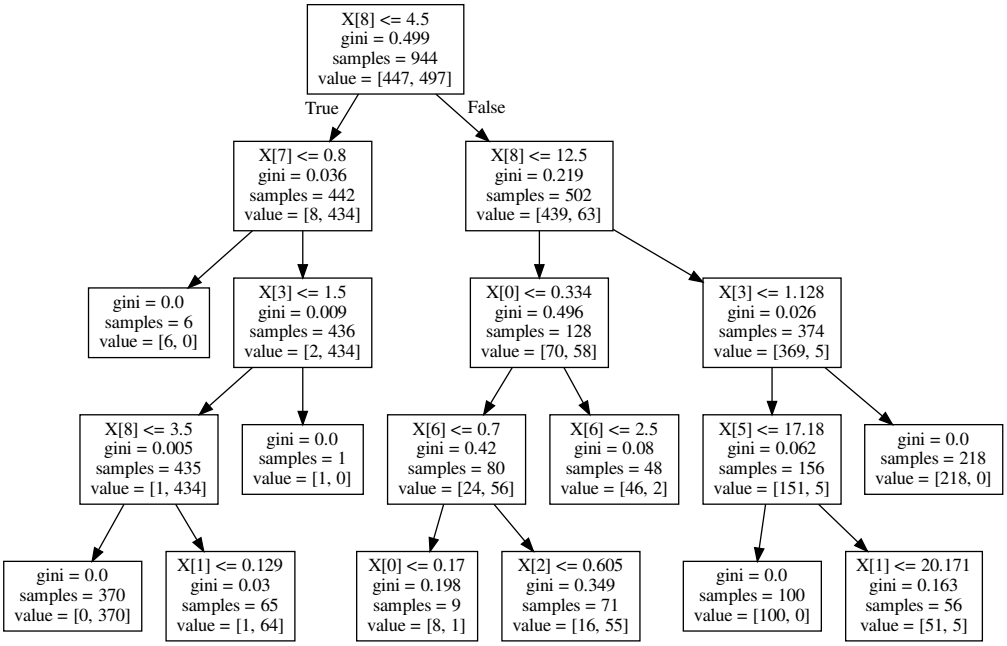
表 4.6 LOO-CV のスコア

指標	スコア
Accuracy	0.960
Precision	0.964
Recall	0.960

表 4.7 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 4.1 構築した決定木 (第 4 段目まで)



4.3.4 考察

本実験の目的である，メタ特徴を用いた特定の属性がカテゴリカル属性であるか否かの判別は，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．

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 [26] 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別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．

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)．例として，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．

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．

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，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．

第 5 章

結論

5.1 まとめ

本研究では，データマイニングの知識を持たないユーザに対しデータマイニングを行うための補助を行うツールを開発するという目標のもと，データマイニングのプロセスのうち前処理に注目し，生データを分析するためにユーザの目的に合わせた前処理を行うアルゴリズムの自動選択に焦点を当ててシステムを開発している．その中で本稿ではカテゴリカル属性をデータマイニングで扱う際の自動前処理方法とデータの性質からカテゴリカル属性であるかを判別する方法を提案することを目的とした．実験の結果，自動前処理については期待通りに動作し，カテゴリカル属性の判別についても訓練データに対する Leave-One-Out Cross-Varidation では 96%，訓練データ外のデータに対してもある程度の正解率が得られ，目的の方法について提案することができた．

本稿では，予めカテゴリカル属性を指定しておくことで，多変量解析において多く用いられているダミー変数化を行うことにより分類データマイニングにおいてもデータを扱えることの確認を行った．その結果，カテゴリカル属性をダミー変数化することによってデータマイニングアルゴリズムの適用は可能であり，数値ラベリングを行った場合と比較して妥当な取扱いが行えていることが確認できた．

また，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する方法の提案も行った．その結果，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．カテゴリカル属性の予測については，第 4 章で述べた通り，データマイニングにおいてカテゴリカル属性として扱うべきか判断がつかない際に，ユーザに提示するというシステムでの活用が考えられるため，適合率を大きく損なわない程度に再現率をさらに向上させることも検討可能である．

しかし，データマイニングの非専門家がデータを解析するために，データの扱いに関する部分も含めて自動化するためには，カテゴリカル属性という大きなくくりでは適切なデータ処理を行えないということが言える．また，本稿では欠損値やエラー値とみなした値を含む事例を全て削除しているため，欠損値やエラー値自体に意味を含んでいる場合や事例の大きな減少に対しては対策がなされていない．

5.2 今後の方針

これらの実験結果を受けて，本研究で今後目指すべき方針をいくつか提示する．

一つは，本研究においてカテゴリカル属性とした，名義尺度や順序尺度をもち間隔尺度や比例尺度を持たない定性的なデータの取扱いについて再考することである．第 4 章第 3 節で示したように，カテゴリカル属性，非カテゴリカル属性という分類では，一様にデータを取り扱うことが適当とは言えない．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検討するべきである．

もう一つは，欠損値やエラー値の取扱いについて考察，検証することである．本稿では，カテゴリカル属性として指定されなかった属性は全て数値として扱い，数値以外が入力されている場合はエラー値として事例ごと削除した．更に欠損値が含まれる事例も削除しているため，前述の通りそれらの値自体に意味があるケースや欠損値やエラー値の多いデータセットには対応できていない．今後はエラー値や欠損値を固有の値として用いる，もしくは類似のデータにより補完するなどのいくつかの方法に対して，それぞれのケースに対応できるような方法について検討すべきである [6, 27]．

また，本稿では自動的に前処理を行うシステムのうちいくつかのエッセンスを提案したが，それらを包括したシステムは構築していない．前述の課題を解決し，システムを構築することも今後行うべきであると考えられる．

謝辞

本研究を進めるにあたり，研究内容やその方針に関するご指導をいただきました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の新美礼彦准教授に心から感謝いたします.

The authors would like to thank Enago (www.enago.jp) for the English language review.

参考文献

[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] D. Dua, C. Graff, “UCI Machine Learning Repository”, http://arhive.ics.uci.edu/ml/ (accessed 2020-01-27), 2019.

[3] V. Vapnik, “The Nature of Statistical Learning Theory”, Springer, 1995.

[4] E. Fix, J.J. Hodges, “Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties”, Technical Report, USAF School of Aviation Medicine, 1951.

[5] E. Fix, J.J. Hodges, “Discriminatory Analysis - Nonparametric Discrimination: Small Sample Performance”, Technical Report, USAF School of Aviation Medicine, 1952.

[6] S. Garcia, J. Luengo, F. Herrera, “Data Preprocessing in Data Mining”, Springer, 2015.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, “The WEKA Data Mining Software: An Update”, SIGKDD Explorations, Vol.11, No.1, pp.10-18, KDD, 2009.

[8] S. Srivastava, “Weka: A Tool for Data Preprocessing, Classification, Ensemble, Clustering and Association Rule Mining”, International Journal of Computer Applications, Vol.88, No.10, pp.26-29, RS Publication, 2014.

[9] “Lightning Fast Data Science Platform for Teams | RapidMiner ©”, https://rapidminer.com/ (accessed 2020-01-27), 2019.

[10] “Amazon SageMaker (機械学習モデルを大規模に構築、トレーニング、デプロイ) | AWS”, https://aws.amazon.com/jp/sagemaker/ (accessed 2020-01-27), 2019.

[11] “KNIME | Open for Innovation”, https://www.knime.com/ (accessed 2020-01-27), 2019.

[12] “Pentaho：日　立”, https://www.hitachi.co.jp/products/it/bigdata/platform/pentaho/ (accessed 2020-01-27), 2019.

[13] A. Filchenkov, A. Pendryak, “Datasets Meta-Feature Description for Recommend-

ing Featureselection Algorithm", Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT), pp.11-18, IEEE, 2015.

[14] G. Wang, Q. Song, H. Sun, X. Zhang, B. Xu, Y. Xhou, "A Feature Subset Selection Algorithm Automatic Recommendation Method", Journal Of Artificial Intelligence Research, Vol.47, pp.1-34, AAAI, 2013.

[15] 南保英孝, 大塚敦史, 木村春彦, 上田芳弘, "メタ特徴の最適化処理による識別器構築アルゴリズム自動選択システム", 科学・技術研究, Vol.5, No.2, pp.179-184, 科学・技術研究会, 2016.

[16] M. Nakamura, A. Otsuka, H. Kimura, "Automatic Selection of Classification Algorithms for Non-Experts Using Meta-Features", China-USA Business Review, Vol.13, No.3, pp.199-205, David Publishing, 2014.

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

[18] A. Gifi, "Nonlinear Multivariate Analysis", Willey, 1990.

[19] 足立浩平, "多変量カテゴリカルデータの数量化と主成分分析", 心理学評論, Vol.43, No.4, pp.487-500, 心理学評論刊行会, 2000.

[20] 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.

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

[22] D. Willshire, G. Kinsella, M. Prior, "Estimating WAIS-R IQ from the National Adult Reading Test: A cross-validation", Journal of Clinical and Experimental Neuropsychology, Vol.13, pp.204-216, Routledge, 1991.

[23] J.R. Quinlan, "C4.5: Programs for Machine Learning", Elsevier, 2014.

[24] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, "Classification and Regression Trees", Chapman and Hall / CRC, 1984.

[25] X. Wu, V. Kumar, "The Top Ten Algorithms in Data Mining", Chapman and Hall / CRC, 2009.

[26] 藤井良宜, "カテゴリカルデータ解析", 共立出版, 2010.

[27] 村山航, "欠損データ分析 -完全情報最尤推定法と多重代入法-", https://koumurayama.com/koujapanese/ (accessed 2020-01-27), 2011.

付録

実験において用いたデータセットの一覧

本節では，本稿で行った実験で用いたデータセットの一覧を提示する．本研究で用いたデータセットは全て UCI Machine Learning Repository より入手した.

実験 1，2

- Absenteeism at work
- Annealing
- Arcene
- Audit Data
- Balance Scale
- Blood Transfusion Service Center
- Breast Cancer Coimbra
- Breast Cancer Wisconsin (Diagnostic)
- Congressional Voting Records
- Cylinder Bands
- Divorce Predictors data set
- gene expression cancer RNA-Seq
- Glass Identification
- Haberman's Survival
- Hayes-Roth
- HCC Survival
- Iris
- Lymphography
- Mammographic Mass
- Primary Tumor
- SPECT Heart, SPECTF Heart
- Tic-Tac-Toe Endgame
- Ultrasonic flowmeter diagnostics {A, B, C, D}
- Vertebral Column {2C, 3C}
- Wine

実験 3

実験 3 においては，実験 1，2 で用いたデータセット群から，カテゴリカル属性と非カテゴリカル属性の数を調整するために，非カテゴリカル属性のみを含み高次元な “Arcene” と “gene expression cancer RNA-Seq” を除き，いくつかのデータセットを追加した.

- 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

実験 2 で用いたメタデータ記述ファイル例

実験 2 で用いたメタデータ記述ファイルの例として，以下に実験の際に作成した "Lymphography" データセットのメタデータ記述ファイルを記載する．

なお，Listing 中の $< datasetdir >$ は，データセットファイルが格納されているディレクトリの絶対参照パスを指定したが，これは同一マシン上のどのディレクトリで動作させても同一の挙動を示すためである．

Listing 1 lymphography.json

```
[{
        "name": "Lymphography",
        "resourcepath": "/<datasetdir>/lymphography.data",
        "filetype": "csv",
        "septype": ",",
        "contain_header": "no",
        "contain_index": "no",
        "no_of_index_column": 0,
        "no_of_target_column": 0,
        "contain_categorical": "yes",
        "categorical_columns":
            [1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18]
}]
```