市民科学における曖昧データを削減するためのデータ収集システム
A Data Collection System with the Function of Reducing Unreliable Data for Citizen Science

盧　宇
LU, Yu

概要：近年，一般市民が科学研究に参加し，研究データ収集作業に協力する市民科学という効率的かつ効果的な研究手法が，環境学や生態学の分野で盛んに応用されている．しかし，市民科学により収集されたデータの中には曖昧なデータが多く存在する．曖昧データの判別を含めて，データの収集，管理，分析などデータに関する作業はほぼ手動で行われているのが日本の市民科学の現状である．本研究では，市民科学における曖昧データを削減するためのデータ収集システムを構築する．本システムは曖昧データの発生を防ぐデータ収集部と防ぎきれなかった曖昧データを除去するデータマイニング部から構成される．データ収集部はデータに関する操作を規範化，自動化しデータマイニング部では決定木を用いて曖昧データを判別する．正解率が高く簡素な決定木を生成する手法 SESAT において，数値型属性に加えカテゴリ型属性をも扱えるように改良する．

Summary: Recently, an efficient research method named Citizen Science is used in environmental and ecological studies. Citizens participate in scientific research and cooperate for collecting study data. But a lot of unreliable data are contained in the data collected by citizens. In the present Citizen Science in Japan, all the works like data collection, management, analyze and the unreliable data discrimination are performed by manual operation. In this research, a data collection system with the function of reducing unreliable data for Citizen Science is developed. The system consists of a data collection part for preventing unreliable data and a data mining part for reducing unreliable data. The former part automates the data collection, the latter part distincts unreliable data with decision tree. SESAT, that is a method for generating simple and accurate decision trees, is improved in order to process not only numerical data but also categorical data.

キーワード: 市民科学・データマイニング・遺伝的アルゴリズム・共生進化・決定木
Keywords: Citizen Science, Data Mining, Genetic Algorithm, Symbiotic Evolution, Decision Tree

1．はじめに

近年，市民科学（Citizen Science）と呼ばれる市民参加型の研究活動が，環境学や生態学の分野で盛んに応用されている[1]．市民科学は，一般市民が科学研究データ収集作業に協力する一方で，研究者が市民を教育することもできる双方向なプログラムである．市民科学を利用することにより研究を効率的かつ効果的に行うことが可能となる．しかし，参加者の専門知識不足のため，収集されたデータには，観測した生物が特定できない“曖昧データ”が混在することが多い．曖昧データの除去は，研究者が手動で行う必要がある．

本研究では，市民科学における曖昧データを削減のためのデータ収集システムを構築する．本システムはデータ収集部とデータマイニング部の 2 つの部分から構成される．データ収集部では，曖昧データの発生を防ぎ，データ収集部で防ぎきれなかった曖昧データをデータマイニング部で除去する．曖昧データの判別には決定木を使用する．正解率の高い簡素な決定木の生成手法として SESAT が提案されているが，数値型属性のデータのみを処理対象としており，カテゴリ型属性のデータを多く含む市民科学に適用することはできない．数値型属性のデータに加え，カテゴリ型属性のデータをも扱えるように SESAT を改良し，本システムのデータマイニング部で利用する．

2．データ収集システムの構成

本システムは曖昧データを削減することを目的として

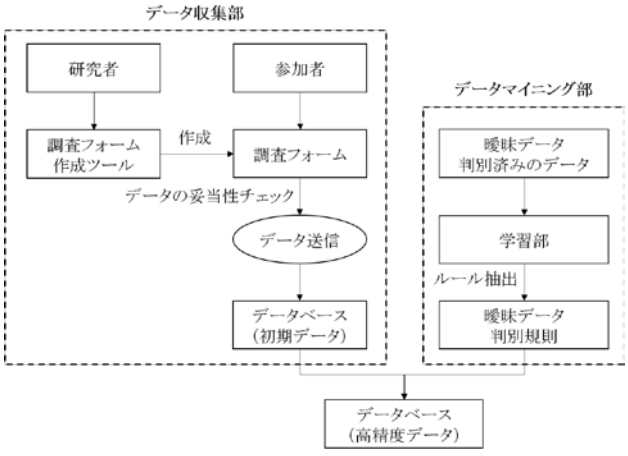

図１．システム構成図

この研究の一部は，情報処理学会第７７回全国大会において発表する予定である．

おり，データの収集・管理・曖昧データチェックの一連の作業を統合的に行える．データを収集するデータ収集部と，曖昧データを判別するデータマイニング部から構成される．システム構成図を図１に示す．

３．データ収集部

データ収集部では，データに関する操作の自動化と人的ミスを防ぐことを目的とし，以下の３つの機能を有する．

- 調査フォーム管理
  調査プロジェクトの具体的な要求に応じて，自由に調査項目を作成，設定することができる．また，作成された調査フォームに対して，編集，削除の操作ができる．完成した調査フォームのサンプルを図２に示す．
- バリデータ
  入力されたデータが仕様に沿って適切に記述されていることを確認し，不適切な箇所があった場合にはエラーとして通知する．
- モジュール管理
  開発者向けの機能であり，新機能を開発する際に，違う機能間互いに影響がないような環境が整っている．

図２．完成した調査フォームのサンプル

４．データマイニング部

４―１．決定木

決定木は木構造を用いて分類規則を表す技法である．決定木（Decision Tree）はノード（葉）とアーク（枝）で構成された木構造である．下に別のノードが接続していないノードを終端ノード，接続しているノードを非終端ノードという．終端ノードにはクラス，非終端ノードには属性，アークには属性値が割り当てられる．非終端ノードの下には，属性値の数だけノードがアークで接続されている．木の根元からスタートし，非終端ノードにある属性と事例の属性値を比較することで，事例を分類できる．

４－２．遺伝的アルゴリズム

遺伝的アルゴリズム（Genetic Algorithm，GA）は，生物の進化のメカニズムを模倣した最適解探索手法である．解空間構造が不明であり，決定的な優れた解法が発見されておらず，または全探索が不可能と考えられるほど広大な解空間を持つ問題に有効である．遺伝的アルゴリズムで扱う問題の解は，遺伝子型（genotype）と表現型（phenotype）の二層構造からなる．遺伝子型は遺伝的アルゴリズムで必要な情報のみにした解の形であり，遺伝的オペレータの操作対象となる．表現型は問題に対する解の形である．問題に応じて表現型から適応度（fitness）を求める．

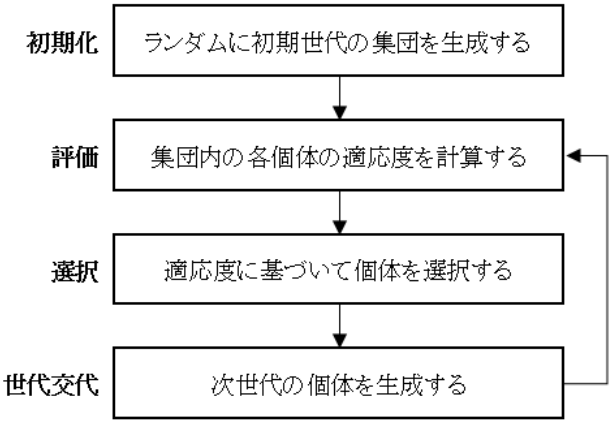

図３．遺伝的アルゴリズムの基本的な流れ

遺伝的アルゴリズムの基本的な流れを図３に示す．遺伝的アルゴリズムでは，解候補を表す個体が複数集まって集団を構成する．各個体は各々遺伝子型として遺伝子コードを持ち，遺伝子型から変換した表現型に応じて適応度が決まる．各個体は世代交代を行い，次の世代の子孫を作り出す．世代交代するときには適応度の良いものほどより多く子孫を作りやすく，適応度の悪いものほど淘汰されやすいようにする．

世代交代する際に，遺伝子型に対して図４に示す交叉，突然変異などの遺伝的オペレータが適用され，次の世代の遺伝子型を生成する．各オペレータの適用頻度，適用部位は一般にランダムに決定される．世代交代によって，次の世代での各個体の表現型が少しずつ変化するとともに，適応度が前の世代よりも良くなる．同様にして，子

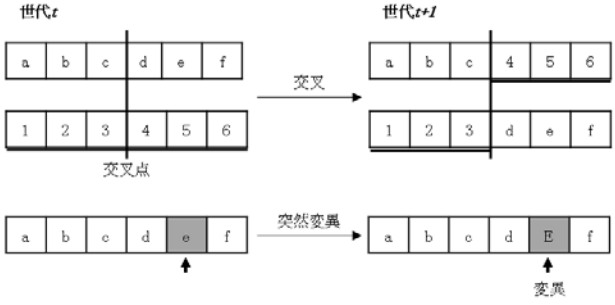

図４．遺伝的アルゴリズムのオペレータ

供の世代が親となって次の世代の子孫を生む．世代交代を繰り返すと次第に集団全体が良くなる．

４－３．共生進化に基づく SESAT

SESAT では，遺伝的アルゴリズムの一種である共生進化に基づいて決定木を生成する[2]．共生進化は，問題の解を複数の部分解に分割し，部分解の組み合わせにより全体解を構成して，部分解集団と全体解集団を並行に進化させることで，最適解を探索する．

決定木を高さ１の部分木（sprig）の組合せと見なし，sprig を部分解，sprig の組合せにより表現される決定木を全体解とする．SESAT の処理手順を図５に示す．

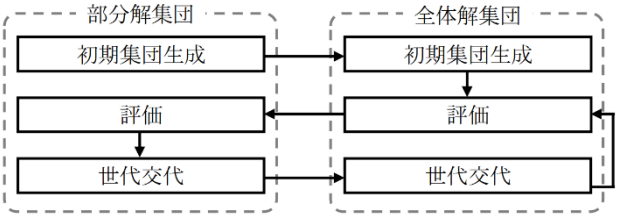

図５．SESAT の処理手順

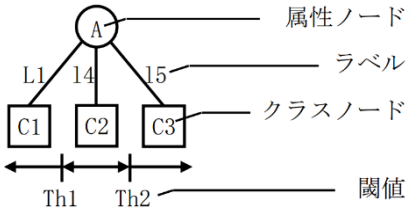

図６．SESAT における sprig の構造

sprig は図６のように１つの属性ノードと複数のクラスノードからなる．クラスノードには，1～C のクラス番号，もしくは 0 が付与される．sprig が決定木に組み込まれる際，クラス番号が付与されたれたクラスノードは終端ノードとなる．また，0 が付与されたクラスノードは非終端ノードとなり，別の sprig の属性ノードが接続される．事例の属性値に従って属性ノードからクラスノードへ走査するときは，アーク間に設定された閾値と属性値の大小関係に従ってたどるアークを選択する．閾値 Th は式１で算出される．

$$Th = (v_{max} - v_{min}) \times \frac{l_a + l_b}{2(M-1)} + v_{min} \qquad \text{(式 1)}$$

l_a，l_b は隣り合う２つのアークのラベル，v_{max}，v_{min} は属性ノードが示す属性の訓練事例における最大値と最小値．M は sprig の最大子ノード数である．

４－４．提案手法 SESAT+

提案手法 SESAT+では，SESAT の sprig の構造，および部分解集団の個体の交叉方法を変更する．

４－４－１．sprig の構造

SESAT+における sprig のクラスノードは，クラス部と属性部から構成される．クラス部では，1～C のクラス番号もしくは 0 を保持する．属性部は，属性ノードが数値型属性を表す場合には使用せず，カテゴリ型属性を表す場合にのみ属性値を格納する．１つの sprig に含まれる

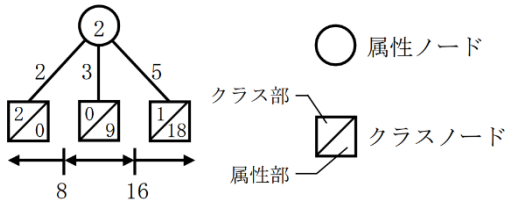

図７．SESAT+における sprig の構造

属性部のうち，必ず１つには 0 が割り当てられ，その他の属性部には互いに重複しないように属性値が割り当てられる．sprig の構造を図７に示す．カテゴリ型属性の事例で決定木を走査するとき，事例の属性値が属性部の値と一致するクラスノードへと進む．一致する属性部がない場合は，属性部の値が 0 のクラスノードへと進む．

カテゴリ型属性の各属性値に数値を割り当てることで，SESAT でもカテゴリ型属性を含むデータを扱うことができる．しかし，決定木の正解率と簡素さは割り当てた数値の大小関係に大きく依存する．正確な分類を目指すと，木が複雑になる可能性もある．提案手法では，クラスノードに属性部を導入することにより，ノード数の増加を抑制している．

４―４―２．クラスノードの染色体表現

クラスノードの染色体は，図８に示すように 32 ビットの２進数で表現する．前半の 16 ビットで表される整数をクラス部に割り当てられた値，後半の 16 ビットで表される整数を属性部に割り当てられた値とする．クラス部と属性部に割り当てられた値は，それぞれ式２と式３を用いて算出される．

$$pType_{Class} = gType_{Class} \% (A+1) \qquad \text{(式 2)}$$

$$pType_{Attr} = gType_{Attr} \% (C+1) \qquad \text{(式 3)}$$

ここで $pType_{Class}$ はクラス部の表現型，$gType_{Class}$ はクラス部の遺伝子型，$pType_{Attr}$ は属性部の表現型，$gType_{Attr}$ は属性部の遺伝型，C はクラス数，A は属性ノードが示すカテゴリ属性の属性値の種類数を表す．

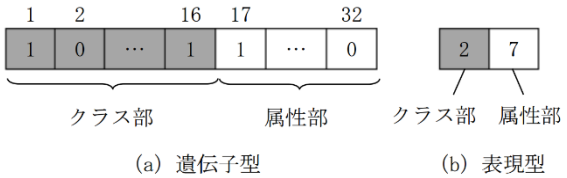

図８．クラスノードの染色体表現

４－４－３．sprig の交叉

SESAT では，属性ノード，およびクラスノードを表す遺伝子の間を交叉点として交叉を行うため，子個体のノードが持つ値は，いずれかの親個体のノードが持つ値となる．提案手法では，染色体の任意のビット間を交叉点として交叉する．親個体に含まれない値も子個体に含まれるようになり，多様な子個体が生成される．

５．評価実験

５－１．データ収集部の評価

被験者が研究者と参加者，それぞれの役を演じて，システムを操作した後，アンケートに回答する．研究者役の被験者は，新しい市民科学プロジェクトを始めることを想定し，初期化されたシステムで，参加者が調査データを送信する環境を整える．まず，システムのサイト名，URL，システム説明，運営者情報などの基本項目を設定し，システムの外観をプロジェクトの目的に応じて変更する．次に，参加者の個人情報記入用フォームを表示する新規参加者登録ページを作成する．最後に，実施する調査ごとに，各調査項目のデータの形式に合わせて，調査フォームを作成し公開する．被験者が上記の一連の作業を完成した後，アンケートに答える．

また，参加者役の被験者は，研究者役の被験者が作成した新規参加者登録ページで，参加者として登録し，システムにログインする．実施中の調査のフォームページで収集した仮データを入力し，送信する．複数の仮データを送信するシミュレーションを繰り返した後で，システムを使用した感想についてヒアリングを受ける．

“各設定項目の説明文が少ないため，設定方法がわからない”という意見により，管理システムには情報系専門知識が必要であることがわかった．情報系の知識を持っていない管理者に優しいヘルプ機能のような管理者支援・教育機能が必要である．システム操作性に対して，本システムでは，クリック，ダブルクリック，右クリック，ドラッグ&ドロップなどのマウス操作を多く取り入れたことにより，画面遷移回数が大幅に減って，操作性が上がった．

データ入力欄セットとバリデータ機能により，人的ミスを防止できるようになったが，さらに重複送信防止機能の追加が求められる．

５－２．データマイニング部の評価

提案手法の有効性を検証するために，UCI 機械学習リポジトリで提供されているデータ agaricus*，statlog，contraceptive method choice(cmc)*，iris に加え，NPO 法人生態教育センターの「お庭の生きもの調査」で収集された市民科学のデータ bird-count*を用いて実験を行った．*の付いているデータはカテゴリ型属性と数値型属性の両方を含んでおり，付いていないデータは数値型属性のみを含んでいる．bird-count*はお庭に現われた鳥についての 2673 件の調査データであり，4 個の数値型属性と 15 個のカテゴリ型属性を含んでいる．sprig の最大子ノード数 M を 5, 3 としたときの 5 分割クロスバリデーションにより得られた正解率とノード数の平均を表 1，表 2 に示す．

表 1：M=5 における正解率とノード数

データ名	SESAT		SESAT+	
	正解率(%)	ノード数	正解率(%)	ノード数
agaricus*	93.95	6.16	93.50	5.46
statlog*	85.68	4.44	85.51	4.00
cmc*	47.75	11.74	47.35	6.68
iris	93.43	4.86	91.60	4.50
bird-count*	75.95	5.24	76.41	3.88
平均	79.35	7.70	78.87	4.90

表 2：M=3 における正解率とノード数

データ名	SESAT		SESAT+	
	正解率(%)	ノード数	正解率(%)	ノード数
agaricus*	97.77	5.18	94.38	5.34
statlog *	85.77	4.76	85.41	3.94
cmc*	46.29	11.44	44.58	5.70
iris	91.83	5.60	91.60	5.32
bird-count*	76.08	6.06	76.42	4.14
平均	79.54	6.61	78.48	4.89

いずれのデータにおいても SESAT より提案手法で簡素な木が生成されており，属性部を導入したことの効果が確認されたといえる．正解率に関しては，SESAT と提案手法で大きな差は見られなかったが，本研究で焦点を当てている市民科学データでは SESAT を上回る正解率が得られている．UCI リポジトリのデータでは，SESAT の正解率を下回ったものの，カテゴリ型属性を含まないデータよりもカテゴリ型属性を含むデータの方が SESAT との正解率の差は小さくなっている．

bird-count*では，再現率が 0.999，適合率が 0.768 となっている．市民科学で収集されたデータは研究に使用されることを考えると，曖昧データをもれなく抽出できることが重要である．したがって，提案手法は市民科学における曖昧データの抽出に関して有用であるといえる．

６．おわりに

本研究では，市民科学の研究データから曖昧データを削減することを目的として，汎用性のあるデータ収集システムを構築した．データ収集システムの構成部分であるのデータ収集部を独立のシステムとして，2014 年 6 月 1 日から NPO 法人生態教育センターの公式サイトで公開され，「お庭の生きもの調査」のデータ収集に活用されている．本システムの導入により，日本の市民科学データ収集作業も，以前のほぼ手作業で行っていた状況が変わり，規範化，自動化される．しかし，現在インターネットが普及している時代でも，市民科学プロジェクトの参加者は高齢者が大半を占めているため，システムの使用率がなかなか上がらない．新システムの使用促進にはまだ工夫が必要である．

また，データマイニング部で提案した SESAT+は，SESAT のクラスノードをクラス部と属性部に分割することにより，数値型属性とカテゴリ型属性の混在するデータの分類を実現している．今後は，生成された決定木を市民の教育に活用することで，収集段階から曖昧データを削減する手法について検討する．

参考文献

1） Janis L. Dickinson， Rick Bonney : Citizen Science: Public Participation in Environmental Research， pp.19-26， Comstock Publishing Associates， 2012

2） 大谷 紀子, 志村 正道：共生進化に基づく簡素な決定木の生成, 人工知能学会論文誌, Vol. 19, No. 5, pp. 399-404, 2004