2024 年度卒業研究概要

太陽光発電機における観測データを用いた教育用データセット生成

大谷 紀子 研究室
2172095　東野　峻大

1. 背景と目的

2022 年 3 月，東京都市大学は，「カーボンニュートラルを実現するための高度デジタル・環境・エネルギー人材育成プログラム」を提案し，文部科学省の「デジタルと専門分野の掛け合わせによる産業 DX をけん引する高度専門人材育成事業」の実施機関に選定された．同年 12 月には，本事業の一環として横浜キャンパスに太陽光発電機等が設置され，実発電量や日射量などが記録されている．記録データを人材育成に活用する方法として，実データを用いた解析演習が挙げられる．解析によりカーボンニュートラルに関する考察が得られた場合，高い学習効果が期待できる．しかし，意義のある結果が得られるデータセットは，解析手法や目的によって異なる．また，時間が経過したデータの使いまわしは学習のモチベーションに影響するため，可能な限り新しいデータであることが望ましい．したがって，目的の解析手法により有意義な結果が得られる最新のデータセットが必要となるが，提供には労力がかかる．

本研究では，特定の解析手法に対して意義のある結果が得られる最新のデータセットを提供することを目的として，解析演習に用いるための実データセットを自動で生成する手法を提案する．対象とする解析手法は，k-means 法と Recurrent Neural Network（以下 RNN）とする．東京都市大学の太陽光発電機を管理するソーラーモニターオフグリッドで取得されたデータを用いて，遺伝的アルゴリズム（GA: Genetic Algorithm）によりデータセットを生成する．

2. k-means 法用のデータセット

染色体は長さ 28 のビット列とし，グレイコードで先頭から 4bit がクラスタ数，9bit がデータセットの先頭日，9bit が末尾日，6bit が乱数シードを表す．乱数シードは，クラスタの初期重心を決めるために用いる．個体を評価する際には，染色体が表すクラスタ数と乱数シードを用いて，k-means 法によるクラスタリングを行う．クラスタリングの対象データは，染色体が表す期間の実発電量，日射量，消費電力量とし，適応度算出には各データに付与されている学内イベント情報を用いる．対象データの数の適切さ DQ と新しさ DN，クラスタリング結果における学内イベントのクラスタ密集度 CD を用いて，個体pの適応度$f(p)$を式(1)で算出する．

$$f(p) = DQ \times DN \times CD \qquad (1)$$

集団サイズは 100，世代交代数は 150 とする．トーナメント選択により親となる 2 個体を決定し，交叉により 2 つの子個体を生成する．交叉確率は 8 割とし，交叉が行われない場合は親個体を次世代に引き継ぐ．加えて，各世代で最も適応度が高い個体を次世代に残すエリート保存戦略を用いる．

2023 年 1 月 1 日から同年 6 月 30 日，12 月 31 日，および 2024 年 5 月 27 日までのデータをそれぞれ 23A，23B，24C とし，提案手法を用いて 3 つのデータからデータセットを生成する．10 回生成したうち最も適応度が高くなった解と解に含まれるイベント数を表 1 に示す．対象とするデータの範囲が変化しても，各最適解が示す約半月の期間には，通常の授業期間に加えて，長期休暇やオリ

表 1　k-means 法用として得られた解

データ	23A	23B	24C
クラスタ数	8	8	8
先頭日	2023/3/29	2023/9/9	2024/4/1
末尾日	2023/4/12	2023/9/23	2024/4/20
乱数シード	32	11	29
イベント数	7	7	7

エンテーションなどが含まれた.

24C で得られた最適解でのクラスタリング結果が考察に適しているかを検証するため，本学 2,3 年生 8 名へインタビューを実施したところ，クラスタ数に関して「少ない方が適切で，クラスタリングの概念を掴みやすい」,「多い方がクラスタの散らばり具合と規則性がわかりやすい」のように，相反する意見が得られた．また，データ分析を授業で扱う本学教員 4 名に同様の結果についてインタビューを実施したところ,「外れ値がない方が演習しやすい」,「時間ではなく日ごとのデータの方が適切」など，教員によって求める課題内容が異なることを示唆する意見が得られた.

3. RNN 用のデータセット

染色体は，長さ 30 のビット列とし，グレイコードで先頭から 9bit が先頭日，9bit が末尾日，5bit が epoch 数，7bit が乱数シードを表す．個体を評価する際には，染色体が表す epoch 数と乱数シードを用いて年，月，日，時刻，日射量，気温，天気から実発電量を求める RNN モデルを生成し，データセットの末尾日から 1 か月間の実発電量を予測する．予測の損失値 VL，処理時間 T，各 epoch 数における精度差分の分散 EV, データセットの新しさ DN，予測値と実測値の一致度を示す決定係数 R^2により，個体pの適応度$f(p)$を式(2)で算出する.

$$f(p) = VL \times T \times DN \times EV / R^2 \qquad (2)$$

集団サイズと世代交代数はともに 15，交叉確率は 7 割とし, 他の条件は k-means 法と同様とする.

提案手法を用いて 23A，23B，24C から 5 回データセットを生成した．最も適応度が高くなった解と解の T，VL，R^2を表 2 に示す．各最適解が示す期間は日数と時期が異なり，2 つのクォーターにまたがる授業期間と約 1 年に渡る期間となった.

表 2　RNN 法用として得られた解

データ	23A	23B	24C
先頭日	2023/4/9	2023/10/13	2023/2/1
末尾日	2023/6/25	2023/11/22	2024/1/4
epoch 数	5	8	5
乱数シード	99	59	27
T	7.731	5.314	27.494
VL	0.013	0.009	0.006
R^2	0.470	0.835	0.717

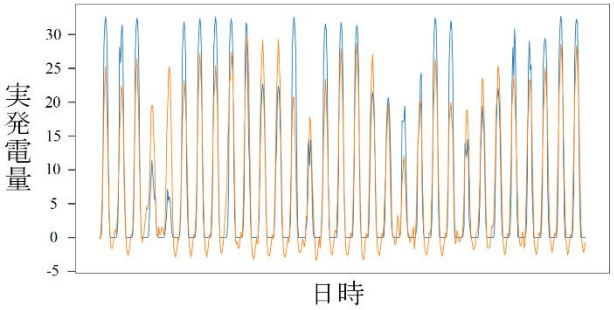

図 1　23B から得られた予測値と実測値

R^2が最も高い 23B の最適解における実発電量の予測値と実測値を図 1 に示す．時刻の変化による実発電量の増減は予測できているが，全体的に予測値は実測値を下回っていた. RNN の予測値を正の値に制限していないため, 負の値も出力される.

4. まとめ

本研究では，k-means 法と RNN の演習に適した最新のデータセットを提供することを目的として，実データセットの生成手法を提案した．k-means 法用の手法では，適度なデータ量で，複数の学内イベントについて考察できるデータセットが生成された．しかし，適切だと感じるクラスタ数は学生によって異なることがわかった．また，各教員にとって適切な課題やデータ形式がさまざまであることがわかったため，生成条件を柔軟に変更できる機能を追加する必要がある. RNN 用の手法では，時系列データの連続的な値の増減を予測できるデータセットを生成できた．今後の課題としては，より適切なパラメータ設定の検討などが挙げられる.