Summary of the 2023 Senior Seminar

Simplified Decision Tree Induction with Multi-objective Symbiotic Evolution

Noriko Otani Laboratory
2072082 Takuya Mitarai

1. Introduction

A decision tree is a method that employs a tree structure to classify data. Non-terminal nodes in a decision tree contain types of attributes and branching conditions, while terminal nodes are assigned classes. The class of unclassified data can be predicted by traversing from the root node to the terminal nodes according to the branching conditions at each node with the attribute values of the data. As decision trees are interpretable models for humans, simplifying the decision tree is crucial.

Symbiotic Evolution is a single-objective genetic algorithm proposed by Moriarty et al [1]. for training the hidden layers of neural networks. Characterized by its divide-and-conquer approach, Symbiotic Evolution simultaneously evolves two populations: partial solutions and complete solutions. The evolution of partial solutions possesses the characteristic of rapid acquisition of effective solutions, while the evolution of complete solutions is characterized by reliable convergence to the optimal solution.

Aiming to generate simplified decision trees, the decision tree generation method SESAT, based on symbiotic evolution, has been proposed [2]. A complete solution, representing a decision tree, consists of several partial solutions representing subtrees of height one. Evolving partial solutions aims to optimize attribute values and branching conditions, while evolving complete solutions focuses on optimizing the tree structure. By operating crossover not only on the tree structure but also on parameters, global search may be feasible. SESAT manipulates a single fitness that is expressed using the accuracy rate and the correct localization rate to minimize the number of nodes. The two rates have a trade-off relationship, and their balance is controlled by the weight α. Only one solution can be obtained by a single search in SESAT. Pareto-optimal solutions, the best trade-offs that satisfy multiple objectives simultaneously, are difficult to enumerate.

This study aims to enumerate optimal decision trees concerning accuracy rate and the number of nodes in a single search. The decision tree induction method NSSESAT (Non-dominated Sorting Symbiotic Evolution for Simple and Accurate Trees) is proposed.

2. Proposed Method

NSGA-II is a representative method of a multi-objective genetic algorithm, and is an algorithm characterized by processes such as the preservation of dominant individuals for the next generation and crowded tournament selection, in addition to the concepts of solution dominance and crowding distance. In the context of solution dominance, a solution A is defined as superior to solution B if it is not inferior to B in all objective functions and is superior in at least one objective function. Crowding distance is a parameter representing the diversity of solutions, indicating that a larger crowding distance corresponds to greater diversity among the solutions.

NSSESAT is an algorithm that incorporates the essence of NSGA-II into SESAT, and is implemented with the same constraints and initialization procedures as SESAT. The procedure for a generation in NSSESAT is as follows.

1. Calculate the fitness of each individual in the complete solution population based on their dominance relationships and crowding distances.
2. Determine the fitness of each individual within the partial solution population based on the fitness of the complete solutions.
3. Preserve and evolve individuals in the partial solution population.
4. Preserve and evolve individuals in the complete solution population.

The fitness $fit(W_i)$ of a complete solution individual W_i is calculated by the equation (1).

$$fit(W_i) = rank(W_i) + \frac{1}{cd(W_i)+1} \qquad (1)$$

where $rank(W_i)$ is the dominance relationship of the accuracy rate and the number of nodes, and $cd(W_i)$ is the crowding distance.

The fitness of a partial solution is the minimum fitness among all complete solutions that refer the partial solution. The top half of the partial solution population is carried over to the next generation as is, while the remaining individuals undergo tournament selection and two-point crossover. Similarly, the top half of the complete solution population is also carried over to the next generation as is, while the remaining individuals are subject to tournament selection and exchange of arbitrary subtrees.

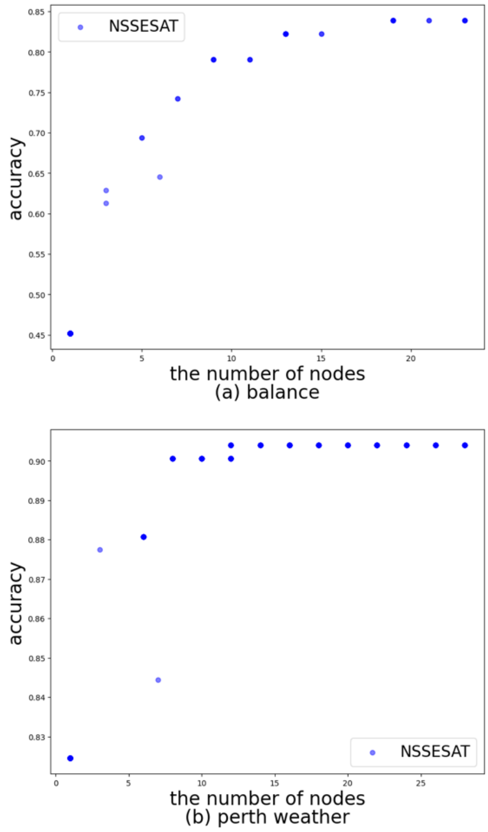
3. Experiments

The experiments were conducted using datasets from the UCI Machine Learning Repository and Kaggle. The characteristics of the datasets are shown in Table 1.

The number of generations was set to 50,000 and 10-fold cross-validation was performed. The Pareto fronts of the execution with the highest number of non-dominated solutions are shown in Figure 1.

Table 1: Characteristics of datasets

Dataset	No. of instances	No. of attributes	No. of classes
balance	625	4	3
perth weather	3025	22	2



(a) balance

(b) perth weather

Figure 1: Pareto front in NSSESAT

4. Discussion

The experimental results show that solutions in a trade-off relationship can be enumerated in a single search by the proposed method. As we can select a preferred decision tree from multiple trees with various accuracy rate and node count according to the objective, NSSESAT may be useful.

References

[1] Moriarty, D.E., Miikkulainen, R.: Efficient Reinforcement Learning through Symbiotic Evolution, Machine Learning, Vol.22, pp.11-32, 1996.

[2] Otani, N., Shimura, M. “Generating the Simple Decision Tree with Symbiotic Evolution,” Trans. JSAI, Vol.19, No.5, pp.399-404, 2004. (in Japanese)