Optimizing Predictions for House Prices using Particle Swarm Optimization

1st Semester of 2024-2025

First Author
emanoil-bogdan.protopopescu@s.unibuc.ro

Second Author
elias.milosi@s.unibuc.ro

Abstract

This project explores the use of ensemble learning combined with Particle Swarm Optimization (PSO) to predict house prices for the Kaggle competition "House Prices - Advanced Regression Techniques." We used three regression models: Random Forest, SVR, and KNeighbors, to generate predictions and subsequently applied PSO to optimize the aggregation of these predictions. Without PSO, our submission achieved a Kaggle score of 0.22084. After incorporating PSO, the score improved significantly to 0.14690, demonstrating the effectiveness of this approach.

1 Introduction

The prediction of house prices is a classical regression problem with practical applications in real estate and finance. The Kaggle competition "House Prices - Advanced Regression Techniques" provides a comprehensive dataset for this purpose.

Previous studies have demonstrated the effectiveness of ensemble methods in real estate price prediction, with Random Forest and Gradient Boosting being particularly popular due to their robustness Breiman (2001); Friedman (2001). Furthermore, Particle Swarm Optimization (PSO) has been successfully applied to optimize ensemble weights in regression tasks Kennedy and Eberhart (1995).

The development of this project involved a collaborative effort during two meetings, where both authors contributed equally to the design, implementation, and evaluation of the proposed method.

This project aims to optimize ensemble predictions using PSO to improve predictive performance. Our contributions include:

- Implementing three regression models: Random Forest, SVR, and KNeighbors, to generate initial predictions.
- Applying PSO to determine the optimal weights for combining predictions from these models.
- Evaluating and comparing the performance of the ensemble with and without PSO optimization.

This approach was chosen to explore how metaheuristic optimization can enhance ensemble learning, a popular method in predictive modeling.

2 Approach

GitHub

2.1 Dataset and Preprocessing

We used the Kaggle dataset "House Prices - Advanced Regression Techniques," which includes various numerical and categorical features. The preprocessing steps involved:

- Removing unnecessary columns (e.g., IDs) and handling missing values by filling them with medians.
- Encoding categorical features using one-hot encoding.
- Standardizing numerical features using `StandardScaler`.

2.2 Regression Models

The following models were used to generate base predictions:

- **Random Forest**: An ensemble method using decision trees, trained with 100 estimators.
- **SVR**: A Support Vector Regressor with an RBF kernel, tuned for optimal C and gamma values.
- **KNeighbors**: A K-nearest neighbors regressor with 7 neighbors.

2.3 PSO Optimization

PSO was used to optimize the weights of the predictions from the three models. The objective function minimized the Mean Squared Error (MSE) on the training data. PSO parameters included:

- **Swarm size**: 100 particles
- **Iterations**: 200
- **Weight bounds**: [0, 1] for each model

2.4 Evaluation

The ensemble prediction without PSO was computed as the mean of individual model predictions. The optimized ensemble combined predictions using weights determined by PSO. Performance was evaluated using the Kaggle scoring metric.

3 Results

- **Without PSO**: The ensemble achieved a score of 0.22084 on Kaggle.
- **With PSO**: The optimized ensemble achieved a significantly improved score of 0.14690.

The results demonstrate the advantage of using PSO for optimizing ensemble models in regression tasks.

4 Limitations

While the PSO-enhanced ensemble improved performance, some limitations were noted:

- **Model Diversity**: The models used had varying performances, with Random Forest dominating the ensemble.
- **Hyperparameter Sensitivity**: PSO results were sensitive to parameter tuning (e.g., swarm size, iteration count).

5 Conclusions and Future Work

This project demonstrated the effectiveness of PSO in improving ensemble predictions for house price estimation. In future work, we aim to:

- Explore other optimization methods (e.g., genetic algorithms).
- Experiment with additional regression models to increase diversity.

References

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

James Kennedy and Russell Eberhart. 1995. Particle swarm optimization. In *Proceedings of ICNN'95-International Conference on Neural Networks*, volume 4, pages 1942–1948. IEEE.