Comparison of Machine Learning Techniques for Land Cover Mapping

Muhammad Hamza Zafar (COSI)

May 16, 2024

1 Introduction

This report evaluates various machine learning methods for land cover classification using multi-seasonal Landsat images and Digital Terrain Models. Techniques assessed include Classification Trees, Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), and Random Forests.

2 Methodology

Each model was implemented using the `mlr` package in R. The models were optimized for hyperparameters through 10-Fold cross-validation, and their performance was evaluated based on mean misclassification error (mmce), accuracy, and kappa statistics.

3 Performance Metrics

In this study, three key performance metrics are used to evaluate the effectiveness of the machine learning models: Mean Misclassification Error (MMCE), Accuracy, and Kappa.

3.1 Mean Misclassification Error (MMCE)

MMCE measures the proportion of incorrect predictions over the total number of cases examined. It is calculated as the number of incorrect predictions divided by the total number of predictions made. A lower MMCE indicates a model with fewer errors, signifying better performance.

3.2 Accuracy

Accuracy is one of the most intuitive performance measures. It is defined as the ratio of correctly predicted observations to the total observations. Higher accuracy indicates a model that better fits the data. It is particularly useful for providing a quick indicative measure of a model's effectiveness but can be misleading in the presence of class imbalance.

3.3 Kappa Statistic

The Kappa Statistic, or Cohen's Kappa, measures the agreement of prediction with the true values. It is normalized by the imbalance of the classes in the data. Kappa values range from -1 (total disagreement) to 1 (perfect agreement), with 0 indicating the agreement that would be expected by random chance. A higher Kappa value suggests that the model has a strong agreement with the true data labels and typically indicates a better model, especially in situations where class distributions are uneven.

3.4 Classification Trees

Used the `rpart` package with hyperparameters optimized as follows:

Parameter	Values
Complexity Parameter (cp)	0.01, 0.05, 0.1
Maximum Depth (maxdepth)	2 to 29
Minimum Split (minsplit)	1 to 50

Table 1: Hyperparameters for Classification Trees

Optimal Parameters: cp=0.01; maxdepth=15; minsplit=23
Performance: mmce=0.2288, accuracy=0.7712, kappa=0.7383

3.5 Random Forest

Random Forest hyperparameter tuning included:

Parameter	Range
Number of Variables (mtry)	1 to 9
Number of Trees (ntree)	1 to 1000

Table 2: Hyperparameters for Random Forest

Optimal Parameters: mtry=5; ntree=889
Performance: mmce=0.1453, accuracy=0.8547, kappa=0.8338

3.6 Artificial Neural Networks

ANNs were tested with varying hidden units, from 1 to 20, each with a decay of 0.01 and maximum iterations of 10,000.

Hidden Units	MMCE	Accuracy	Kappa
1	0.738	0.262	0.147
2	0.613	0.387	0.295
⋮	⋮	⋮	⋮
12	0.186	0.814	0.787
20	0.208	0.792	0.761

Table 3: Performance of ANNs with Different Hidden Units. The best result is achieved with 12 hidden units, showing significant accuracy and kappa scores.

3.7 Support Vector Machines

SVMs were configured with various kernels and parameters:

Parameter	Configuration
Kernel	vanilladot, rbfdot, polydot, tanhdot
Cost (C)	Exp. scale from log(0.1) to log(100)
Gamma (for rbfdot)	Exp. scale from log(0.05) to log(1)
Degree (for polydot)	1 to 10
Scale (for polydot and tanhdot)	1 to 10

Table 4: Hyperparameters for Support Vector Machines

Optimal Parameters: kernel=vanilladot; C=6.24
Performance: mmce=0.1032, accuracy=0.8968, kappa=0.8820

4 Summary of Best Model Performances

A comparative overview of the best-performing models across different techniques.

Model	MMCE	Accuracy (%)	Kappa
Support Vector Machines	0.103	89.7	0.882
Random Forest	0.145	85.5	0.834
Artificial Neural Networks	0.186	81.4	0.787
Classification Trees	0.229	77.1	0.738

Table 5: Comparison of the best model performances from each technique, sorted by accuracy. SVM shows the highest accuracy and kappa, closely followed by the Random Forest.

5 Conclusion

The comparison of machine learning techniques revealed diverse strengths across models, with SVM showing the highest accuracy and kappa values, suggesting superior performance for this dataset and settings.

6 Detailed Model Analysis using Random Forest

The results discussed were obtained from the following outputs of the Random Forest model:

- Confusion Matrix: `model.rf$learner.model$confusion`
- Variable Importance: `model.rf$learner.model$importance`

6.1 Class Confusion and Accuracy

The confusion matrix provides a detailed look at how well each class is predicted against the others, highlighting instances of misclassification.

	Predicted Class							
True Class	1	2	3	4	5	6	7	8
1	46	0	1	1	0	0	0	2
2	0	45	0	0	0	0	5	0
3	0	0	40	0	0	0	4	6
4	0	0	0	40	8	0	0	2
5	0	0	0	4	45	0	0	2
6	0	0	1	0	0	48	0	0
7	0	9	2	0	0	1	40	5
8	1	0	4	0	2	0	2	41

Table 6: Confusion Matrix of the Classification Model

Analysis:

- **Best Ranked Class:** Class 6 has the lowest class error, indicating it is the best predicted.
- **Worst Ranked Class:** Class 7 has the highest class error, making it the worst predicted.
- Classes 3 and 7 show significant confusion with Class 8, suggesting spectral or feature overlap.

6.2 Variable Importance

Analysis of variable importance helps in identifying which features are most influential in the model.

Variable	Mean Decrease in Gini
BlueV	13.266
GreenV	11.683
RedV	15.461
RedEdge1V	11.889
⋮	⋮
SWIR2P	24.426

Table 7: Variable Importance Measured by Mean Decrease in Gini

Most Important Variables:

- The variables *RedP* and *SWIR2P* show the highest mean decrease in Gini, indicating their critical role in the model's decision-making process.

Differences in Importance Metrics:

- *Mean decrease in accuracy* focuses on the direct impact of each variable on the model's predictive accuracy when omitted.
- *Mean decrease in Gini* considers how much each variable contributes to the homogeneity of the nodes and leaves in the model.
- Typically, variables important for accuracy also reduce Gini impurity significantly, but this isn't always the case, reflecting different roles variables may play in prediction versus classification structure.

References

[1] Rodriguez-Galiano, V.F. and M. Chica-Rivas, "Evaluation of different machine learning methods for land cover mapping of a Mediterranean area using multi-seasonal Landsat images and Digital Terrain Models." *International Journal of Digital Earth*, 2014, 7(6): 492-509.