A COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR MULTI-CLASS CLASSIFICATION: DECISION TREES, RANDOM FORESTS, NAIVE BAYES, SVM, KNN

Uchenna Chima
Department of Computer Science
Dalhousie University
Halifax, Canada
B00949727
https://github.com/uchechim/CSCI6515_ML_Project

Abstract

In my research study, I conduct a comparative analysis of five different machine learning algorithms which are applied to three different datasets to solve multi-class classification problems. I evaluate the performance of Decision Trees, Random Forests, Naive Bayes, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) using a variety of heuristic evaluation metrics on the following datasets: Wine Quality, Nursery, and Covertype. The purpose of using a variety of datasets within different domains was to see how each machine learning algorithm performs when given a different type of dataset to generate a holistic solution. The Wine Quality dataset is a small dataset that contains 4,898 instances, the Nursery dataset is a large dataset that contains 12,960 instances and the Covertype dataset is the largest and contains 581,012 instances. The results from the study show strengths and weaknesses among these algorithms thus providing further insight into their robustness in the context of multi-class classification tasks. Although the results from the research conducted are in favor of the Random Forest algorithm as it received the highest accuracy and the best balance between other heuristic evaluation metrics, this study provides valuable insights for machine learning enthusiasts and researchers to assist them in selecting the most appropriate algorithm for their specific machine learning task.

Keywords: Machine Learning, SVM, KNN, Decision Tree, Random Forest, Naive Bayes

1. Introduction

In the rapidly evolving field of technology, machine learning has been a subject of tremendous growth and expansion over the last few decades. The use of machine learning in modern applications has grown significantly and with this growth, the need to develop robust algorithms to meet consumer's demands has not slowed down. In machine learning, one of the critical tasks is multi-class classification which entails training a machine learning model to be able to classify instances that belong to three or more classes or categories. In my research, I dive into the realm of applying five different machine learning algorithms (Decision Trees, Random Forests, Naive Bayes, SVM, and KNN) to answer the question: "Which machine learning algorithm(s) perform the best when presented multi-class classification tasks?" To answer the proposed question, I conducted a comprehensive comparative analysis that utilizes various heuristic evaluation metrics in addition to visualizations to gain further insight into the data that was used to conduct the study for the proposed question. The purpose of this study is to provide valuable insights regarding which algorithm to potentially select when given a multi-class classification task.

2. Literature Review

After reviewing a plethora of relevant research studies that relate to my topic of discussion, it was revealed that previous work done in the field has explored the strengths and weaknesses of individual algorithms in achieving optimal classification performance regarding multi-class classification.

In a notable study conducted by Gursev Pirge [2], he conducted a comparative analysis on a relatively small dataset that consisted of only 62 samples with class labels ranging from 1-4 to represent the different types of metals. Amongst the six algorithms that were compared, he utilized four of which I have chosen to do my study on which are: Decision Trees, Random Forest, Naive Bayes, and SVM. His study results showed that the Random Forest classifier achieved the most optimal performance as boasted a 94.7% accuracy with the Decision Tree algorithm falling just a little shy of that score.

In a study conducted by Lampe et. al [3], a comparative analysis was conducted in which two of the algorithms which I have chosen to do my study were assessed: Random Forest and SVM out of the four total algorithms chosen in this study. Unlike my study and the study conducted by Gursev Pirge [2], Lampe et al. wanted to see how four different algorithms performed on image data. It was found that the deep forward neural network outperformed and yielded the best results when compared to the other algorithms for classifying multi-class image data.

Building upon the previous work that has been done in the context of algorithm selection for multi-class classification tasks, my study seeks to provide a comprehensive and up-to-date assessment of the aforementioned algorithm's performances.

3. Methods and Experiments

My research employs a systematic approach to evaluate the performance of Decision Trees, Random Forests, Naive Bayes, SVM, and KNN. My research began by selecting three datasets that are comprised of different qualities. The names

of the three datasets used were: Wine Quality, Nursery, and Covertype. Each one of these datasets differed in regards to size(number of instances), number of features, and number of categories present in the multi-class label.

The Wine Quality dataset is a small dataset that contains 4,898 instances, 11 features, and 11 different categories (sensory data) in the multi-class label which was further discretized into three distinct categories. The discretization process involved categorizing the original given labels into three categories to represent the quality of the wine. For this dataset, instances in the range of 1-5 were discretized as "0" to represent low-quality wine, instances in the range of 6-7 were discretized as "1" to represent average-quality wine, and instances in the range of 8-10 was discretized as "2" to represent high-quality wines. The goal for this dataset was to see how well the aforementioned machine learning algorithms did in classifying the wine qualities based on the 11 different physiochemical properties.

The Nursery dataset is a large dataset that contains 12,960 instances, 8 features, and 5 different categories (admission recommendation) in the multi-class label. Although discretization was not performed on this dataset, label encoding was required as the original dataset provided discrete values.

The Covertype dataset is the largest of all three datasets and contains 581,012 instances, 52 features, and 7 different categories (forest cover types) within the multi-class label. Unlike the Nursery dataset and the Wine Quality Dataset, no further modifications to the multi-class labels were required.

1. Pre-Processing

Every dataset underwent a similar pre-processing procedure that consisted of: Data preparation, Data cleaning, Data transformation, Outlier removal Oversampling (if required), and Data splitting. The data preparation phase involved loading the data and performing any necessary steps to ensure that the data frame was ready to proceed to the next pre-processing step. For the Wine Quality Dataset, a concatenation of the individual Red and White wine NumPy files was required in addition to ensuring that the columns were aligned properly. For the Nursery Dataset and CoverType dataset, I prepared the data by providing columns with their appropriate labels. All three datasets underwent the same data cleaning procedure and transformation process which involved removing duplicates and null values and normalizing/standardizing features based on their nature (continuous vs discrete). All three datasets utilized the z-score method for outlier removal, and all datasets except the CoverType dataset required oversampling. For the Wine Quality and Nursery Datasets, a train/test/validation split of 60/20/20 was used as hyperparameter optimization using GridSearchCV was applied. For the CoverType dataset, a train/test split of 70/30 was applied.

2. Model Architecture

For each dataset, the: Decision Tree, Random Forest, Naive Bayes, SVM, and KNN algorithms were applied. For the Wine Quality and Nursery datasets, each model excluding Naive Bayes utilized hyperparameter optimization using GridSearchCV with a K-fold value of 5. For each model, the best parameters were deduced using their respective hyperparameter search space. The hyperparameter search space for the decision tree model was: Criterions of Entropy vs Gini with

varying max_depths of 0, 10, 20, and 30. The hyperparameter search space for the random forest model was: N_estimators of 50, 100, and 200 with varying max_depths of 0, 10, 20, and 30. The hyperparameter search space for the SVM model was: C values of 0.1, 1, 10 with varying kernels of 'Linear', 'RBF', and 'Poly'. The hyperparameter search space for the KNN model was: N_neighbors of 3, 5, and 7 with varying weight values of 'uniform' and 'distance'. Due to resource and time constraints, the largest dataset (cover type) did not utilize hyperparameter optimization. For this dataset, the models were architected using the best hyperparameters deduced from the second largest.

3. Heuristic Evaluation Methods

For each dataset, the: Decision Tree, Random Forest, Naive Bayes, SVM, and KNN algorithms were trained and their performance was evaluated using the same methods and metrics. The evaluation methods used for each algorithm to measure effectiveness are as follows: Confusion Matrices, Accuracy, Precision, Recall, and F1_Score.

4. Results

The results generated from applying the various heuristics on three different datasets show compelling insights into the performance of each algorithm. All in all, the Random Forest algorithm boasted the highest accuracy score and the best balance between precision and recall scores. The statistical metrics provided below are based on the previously mentioned evaluation metrics and have been shown to represent the result of each heuristic on each dataset.

	Wine Quality Dataset				
	Decision Tree	**Random Forest**	**Naive Bayes**	**SVM**	**KNN**
Accuracy	72.20%	80.00%	49.90%	69.60%	73.00%
Precison	75.90%	81.70%	56.20%	71.00%	76.70%
Recall	72.40%	80.60%	50.70%	68.90%	72.30%
F1 Score	73.80%	81.10%	50.80%	69.80%	73.90%

Table 1. Wine Quality Dataset Heuristic Results

		Nursery Dataset			
	Decision Tree	**Random Forest**	**Naive Bayes**	**SVM**	**KNN**
Accuracy	99.40%	98.30%	67.40%	96.90%	87.30%
Precison	99.40%	98.40%	77.30%	97.70%	90.80%
Recall	99.40%	98.80%	77.60%	96.80%	87.30%
F1 Score	99.40%	98.50%	69.70%	97.30%	88.60%

Table 2. Nursery Dataset Heuristic Results

	Covertype Dataset				
	Decision Tree	**Random Forest**	**Naive Bayes**	**SVM**	**KNN**
Accuracy	93.80%	94.90%	13.60%	72.40%	94.60%
Precison	84.20%	80.90%	45.90%	33.90%	83.40%
Recall	84.50%	92.60%	25.40%	34.10%	87.60%
F1 Score	84.30%	85.30%	9.70%	32.00%	85.30%

Table 3. Covertype Dataset Heuristic Results

5. Discussions

When performing an analysis on the Wine Quality dataset, although Random Forest exhibits the highest accuracy score (80.00%), precision score (81.70%), recall score (80.60%), and F1_Score (81.10%) and is the top-performing algorithm on this dataset, Decision Tree also did quite well though it lagged slightly in terms of accuracy (72.20%) and F1_Score (73.80%). The K-NN algorithm also was quite competitive as it bolstered an accuracy score of (73.00%) and also had a good balance between precision and recall (76.70% & 72.30%) respectively. SVM was a little behind Decision Tree and KNN in regards to accuracy and F-Measure but still offered a reasonable performance. Last but not least, Naive Bayes performed the worst among all five algorithms with an accuracy of (49.90%), and precision, recall, and F1 scores of (56.20%, 50.70%, and 50.80%) respectively. After analyzing the results on this dataset (small dataset - 4,898 instances), the Random Forest algorithm would be the one I would recommend using as it outperformed the other four algorithms across all metrics.

Unlike the small dataset (wine dataset), when considering the five Machine Learning algorithms used in my experiments for the large dataset (nursery dataset - 12,960 instances), the results varied quite a bit. On this dataset, the Decision Tree and Random Forest algorithm exhibited near-identical results with accuracy, precision, recall, and F1_Scores above 98.00% thus showing a near-perfect classification. Unlike the Wine Quality dataset, KNN slightly underperformed when compared to SVM as it had an accuracy score of (87.30%) whereas SVM had an accuracy score of (96.90%). Both KNN and SVM had balanced precision and recall scores. Last but not least, the least desired algorithm was Naive Bayes as it scored an accuracy of 67.40% and had a significantly lower precision, recall, and F1_score in comparison to the other four algorithms. When providing a recommendation for a slightly larger dataset, a choice between using the Decision Tree or Random Forest heuristics would suffice as they both achieved near-perfect classification results. However, the Decision Tree algorithm had a very slight advantage over Random Forest in terms of precision, recall, and F-score with an average of about 1.00%.

Unlike the small and large datasets (wine dataset, nursery dataset), when considering the five Machine Learning algorithms used in my experiments on the largest dataset(cover type dataset - 581,012 instances), the results varied but were more similar to the results achieved on the nursery dataset. On this dataset, the Random Forest algorithm had the highest accuracy and F-Scores with scores of

94.90% and 85.30% respectively. In comparison, the Decision Tree algorithm had an accuracy and F-Scores of 93.80% and 84.30% respectively. However, on this dataset, Random Forest scored 80.90% on precision and 92.60% on recall whereas Decision Tree scored 84.20% on precision and 84.50% on recall. KNN was quite competitive in comparison to Decision Tree and Random Forest as it yielded accuracy, precision, recall, and F-Scores of 94.60%, 83.40%, 87.60%, and 85.30%. Unlike the previous two datasets, the SVM algorithm had a significantly lower performance as did the Naive Bayes algorithm. SVM bolstered a 72.3% accuracy but scored less than 35% on precision, recall and F1 Scores and Naive Bayes showed a 13.6% accuracy with a highly fluctuating precision, recall, and F1 Score thus making these two the least desirable algorithms for this dataset.

When providing a recommendation, based on the metrics extracted from all three datasets, Random Forest would be the most desirable as it shows the highest accuracy and the best balance between precision, recall, and F1 scores.

6. Conclusion

In conclusion, my research sheds light on the importance of algorithm selection in achieving optimal classification performance in the case of multi-class classification problems. The Random Forest algorithm emerged as the best choice while the Naive Bayes proved to be quite suboptimal. SVM and KNN show promising potential but as with most machine learning tasks, knowing the domain and scope of your problem may give you a little more insight if you prefer to choose either one of these algorithms. Further research could explore different types of algorithms and dataset types in addition to addressing dataset biases more in-depth than has been done in my study to further advance the problem of algorithm selection when considering multi-class classification tasks.

References

1. Pirge, G. (2020, December 28). *Performance comparison of multi-class classification algorithms*. Medium. https://gursev-pirge.medium.com/performance-comparison-of-multi-class-classification-algorithms-606e8ba4e0ee

2. Nitze, I., Schulthess, U., & Asche, H. (2012, May). Comparison Of Machine Learning Algorithms Random Forest, Artificial Neural Network, and Support Vector Machine to Maximum Likelihood for Supervised Crop Type Classification. https://www.researchgate.net/profile/Ingmar-Nitze/publication/275641579_COMPARISON_OF_MACHINE_LEARNING_ALGORITHMS_RANDOM_FOREST_ARTIFICIAL_NEURAL_NETWORK_AND_SUPPORT_VECTOR_MACHINE_TO_MAXIMUM_LIKELIHOOD_FOR_SUPERVISED_CROP_TYPE_CLASSIFICATION/links/5541238e0cf2b790436bc791/COMPARISON-OF-MACHINE-LEARNING-ALGORITHMS-RANDOM-FOREST-ARTIFICIAL-NEURAL-NETWORK-AND-SUPPORT-VECTOR-MACHINE-TO-MAXIMUM-LIKELIHOOD-FOR-SUPERVISED-CROP-TYPE-CLASSIFICATION.pdf

3. Lampe, L., Niehaus, S., Huppertz, HJ. et al. Comparative analysis of machine learning algorithms for multi-syndrome classification of neurodegenerative syndromes. Alz Res Therapy 14, 62 (2022). https://doi.org/10.1186/s13195-022-00983-z