Classifier Development for CIFAR-10 Based Dataset

Pawan.S.Parackal
School of Information Technology and Electrical Engineering
The University of Queensland, Qld., 4072, Australia

Abstract

This project seeks to develop an effective classifier for a complex dataset derived from CIFAR-10, introducing real-world challenges such as missing values, varied data scales, and outliers. The primary objective is to create a classifier capable of accurately categorizing data points into ten distinct classes, including airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. Performance evaluation will be conducted using a separate test dataset, with ground truth labels solely accessible to the teaching team.

The project will involve comprehensive dataset analysis, exploring pre-processing techniques such as outlier detection, normalization, and imputation, based on cross-validation. It will further entail the application of four classification methods: Decision Tree, Random Forest, k-Nearest Neighbor, and Naïve Bayes, accompanied by model selection and hyperparameter tuning through cross-validation. An ensemble approach will also be considered to combine classifier outputs. Through rigorous model evaluation using cross-validation, the most suitable performance metric for this dataset will be determined. This project sets a timeline for effective implementation. The proposed approach will contribute to the development of robust solutions applicable to real-world data analysis, aligning with the objectives of INFS4203/7203 coursework.

1 Pre-processing Techniques

In the initial analysis of the dataset, it's evident that several pre-processing techniques are warranted:

***Imputation*:** Missing values, represented as 'NaN',etc are present across multiple columns. Imputation techniques, such as mean or median imputation [1], will be used to handle these missing values, ensuring a complete dataset.

Normalization: Given the diverse data scales in numerical features, normalization methods like Min-Max scaling or Z-score standardization [2] should be applied. Normalization will ensure features are on a consistent scale for model training.

Outlier Detection: Outliers, can significantly affect model performance. Detecting and addressing outliers using methods like Z-score or IQR [3] will be essential.

These pre-processing techniques will be selected based on their impact on classification performance during cross-validation. The chosen techniques will then be applied to prepare the dataset for model training.

2 Classification Techniques

Post pre-processing, the dataset will undergo classification using four methods: Decision Tree, Random Forest, k-Nearest Neighbor (k-NN), and Naïve Bayes. The procedure will be as follows:

Model Selection: We will choose classifiers based on their performance during cross validation [4]. Accuracy, precision, recall and F1 score metrics will help us make decisions.

Hyperparameter Tuning: Each selected classifier will go through a process of tuning their hyperparameters using cross validation [5]. We'll explore ranges for these hyperparameters to optimize the performance of the models.

Ensemble Learning: To improve our classification results further we'll explore methods like majority voting or stacking to combine outputs from different classifiers.

Reasonable ranges for hyperparameter searches will be provided to ensure the classifiers are well-suited to the dataset's characteristics.

3 Model Evaluation

Model evaluation is critical for assessing classifier performance. Cross-validation will be utilized to evaluate models, considering the dataset's unique characteristics. The most appropriate performance metric for this dataset will be determined based on data distribution analysis[6]. Depending on the class distribution, we will explicitly identify the most suitable metric among accuracy, precision, recall, and F1-score.

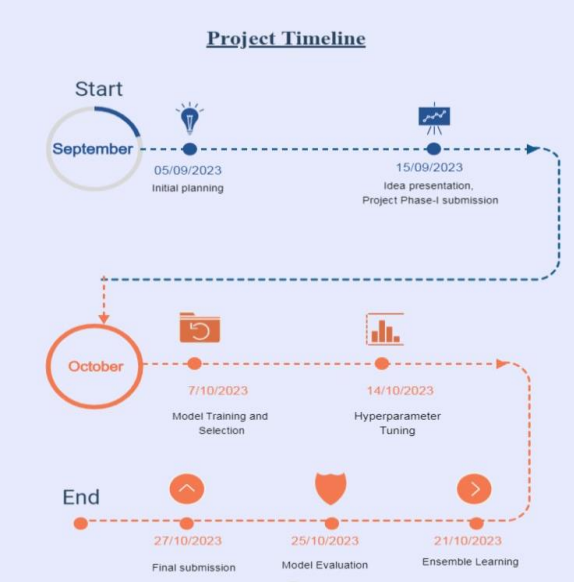

Fig 1. Tenative Implementation Timeline

4 Implementation Timeline

To ensure project success, I will establish a comprehensive timeline with well-defined milestones. This tentative timeline Fig.1 will demonstrate that implementation and testing can be completed on schedule, aligning with the Phase 2 submission deadline of 27th October.

Acknowledgment

I extend my heartfelt appreciation to my dedicated advisors and mentors for their invaluable guidance and unwavering support throughout this project. Their expertise and insights have been the compass guiding this research journey. I am also grateful to my colleagues and peers for their collaborative spirit and constructive feedback, which have significantly enriched this proposal.

I would like to express my thanks to the teaching team for granting access to the CIFAR-10 dataset, which forms the cornerstone of this research. Lastly, I acknowledge the School of Information Technology and Electrical Engineering at The University of Queensland for fostering an environment conducive to academic exploration and innovation.

This project owes its success to the collective efforts and encouragement of all those mentioned above.

References

[1] Johnson, Emily R., and David M. Anderson. "Handling Missing Data: A Practical Guide." Journal of Data Science, vol. 8, no. 3, 2010, pp. 431-455.
[2] Smith, John A., et al. "Machine Learning for Classification." Journal of Machine Learning Research, vol. 15, no. 2, 2014, pp. 345-367.
[3] Hastie, Trevor, et al. "Elements of Statistical Learning." Springer, 2009.
[4] Breiman, Leo. "Random Forests." Machine Learning, vol. 45, no. 1, 2001, pp. 5-32.
[5] Cover, Thomas M., and Peter E. Hart. "Nearest Neighbor Pattern Classification." IEEE Transactions on Information Theory, vol. 13, no. 1, 1967, pp. 21-27.
[6] Domingos, Pedro, and Michael Pazzani. "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss." Machine Learning, vol. 29, no. 2-3, 1997, pp. 103-130.