Comparing Supervised Learning Algorithms

Cecilia Truong
COGS 118A
13 December 2024

Abstract

This work investigates the performance of three supervised learning algorithms (SVM with RBF kernel, KNN, and Decision Tree) on three datasets: Iris, Wine, and Breast Cancer (Wisconsin). The datasets were preprocessed and partitioned into three training and testing splits (80/20, 50/50, and 20/80) to evaluate the classifiers' accuracy under varying conditions. The results demonstrate that SVM consistently performed with the highest accuracy across all datasets. Decision Tree performed closely, often nearing SVM's performance, while KNN lagged behind, especially on datasets with greater complexity. These findings align with the trends observed in the empirical analysis by Caruana and Niculescu-Mizil, but refinements in the methodology to address overfitting could improve alignment with prior research.

1 Introduction

This paper investigates the accuracy performances of three supervised learning classifiers (**SVM with RBF kernel, KNN, and Decision Tree**) across three well-known datasets sourced from UCI's repository: **Iris, Wine, and Breast Cancer Wisconsin (Diagnostic)**. This study is inspired by Caruana and Niculescu-Mizil's work on the empirical analysis of supervised learning algorithms, which evaluated a wide array of models on multiple datasets. While their work was extensive, this report applies the core ideas on a smaller scale, focusing on three classifiers and three datasets. By doing so, it aims to provide an in-depth application of the concepts learned in COGS 118A and examine the comparative performance of these classifiers. The goal of this paper is to align with the findings of Caruana and Niculescu-Mizil while offering a concise and focused investigation. It seeks to analyze the accuracy of the classifiers under consistent experimental conditions, providing information on their relative strengths and weaknesses.

2 Problem & Data Description

The primary goal of this paper is to evaluate and compare the performance of three supervised learning classifiers on selected datasets. This study aims to align with the findings of Caruana and Niculescu-Mizil's empirical analysis of supervised learning algorithms, with some expected deviations due to differences in dataset selection and experimental setup. While the datasets are sourced using scikit-learn's library, they correspond directly to the versions available in the UCI Machine Learning Repository.

2.1 Iris Dataset

The Iris dataset is a simple, tabular dataset frequently used throughout COGS 118A. It contains 150 instances and 4 continuous features: sepal length, sepal width, petal length, and petal width. Each instance represents a plant, and the target variable classifies it into one of three species: Iris Setosa, Iris Versicolour, or Iris Virginica. Due to its simplicity and well-separated classes, this dataset serves as a baseline for evaluating the classifiers.

2.2 Wine Dataset

The Wine dataset includes chemical analysis results of wines grown in the same region of Italy but derived from three different cultivars. It comprises 178 instances and 13 continuous features, such as alcohol content, flavanoids, and color intensity. The target variable categorizes each instance into one of three wine classes. This dataset introduces higher dimensionality and slightly imbalanced classes, making it more challenging than the Iris dataset.

2.3 Breast Cancer (Wisconsin) Dataset

The Breast Cancer dataset is a multivariate dataset used to classify breast masses as malignant or benign based on characteristics of cell nuclei extracted from digitized images. It contains 569 instances and 30 features. These 30 features are derived from 10 unique measurements (radius, texture, perimeter, area, compactness, etc.), with each measurement represented as the mean value, standard error, and worst (maximum) value. The target variable represents one of two diagnoses: malignant or benign. This structure provides a comprehensive view of the cell nuclei's characteristics and is the most complex out of the three datasets.

3 Methodology

Before implementing the classifiers, the datasets were preprocessed to simplify their class structures. For the Wine and Breast Cancer datasets, which originally had more than two classes, the classes were merged into two groups. Afterward, each dataset was divided into three distinct training and testing splits: 80/20, 50/50, and 20/80.

3.1 Classifiers

SVM with RBF Kernel: The Support Vector Machine classifier uses a radial basis function kernel to model non-linear relationships in the data. The hyperparameter gamma values ranged from 10^{-6} to 10^{-2}, and C values ranged from 1 to 10,000. For each combination of C and gamma, 3-fold cross-validation was used to calculate average validation accuracy and training error. The combination that yielded the highest validation accuracy was selected as the optimal configuration.

KNN: The KNN classifier predicts the class of an instance based on the majority vote/average of its nearest neighbors. The number of neighbors, *k*, is the primary hyperparameter and was optimized by evaluating three values: *k=1,2,3*. For each *k*, 3-fold cross-validation was used to calculate the average validation accuracy and training error. The optimal *k* was selected based on the lowest average validation error.

Decision Tree: The Decision Tree classifier's key hyperparameter, max_depth, was optimized through a grid search conducted over the values [1, 2, 3, 4, 5] using 3-fold cross-validation. The training and validation accuracies for each depth were recorded, and the most optimal max_depth was selected based on the highest validation accuracy.

4 Experiments

Iris Dataset: On the Iris dataset, SVM achieved the highest accuracy of 0.9867, followed by Decision Tree at 0.9667 and KNN at 0.9467. The dataset's simplicity and well-separated classes made it favorable for SVM, which excels at creating smooth decision boundaries. Decision Tree also performed well, capturing the dataset's structure effectively. KNN performed slightly worse, but nonetheless on par with the rest of the classifiers.

Breast Cancer(Wisconsin) Dataset: Both SVM and Decision Tree achieved the highest accuracy of 0.9561 on the Breast Cancer dataset, significantly outperforming KNN, which reached 0.8421. This dataset's high dimensionality highlighted SVM's ability to handle complex feature interactions and Decision Tree's capacity to model patterns effectively. KNN struggled in performance most likely due to its incompatibility with complexity, as its distance-based approach becomes less reliable with a large number of features.

Wine Dataset: On the Wine dataset, SVM again outperformed the other classifiers with an accuracy of 0.9775, followed closely by Decision Tree at 0.9722. KNN had an accuracy of 0.8056, falling behind the other two classifiers. Like in the Breast Cancer results, the greater feature complexity and slight class imbalance of the Wine dataset demonstrated SVM's and Decision Tree's flexibility, while KNN's struggles were consistent with its sensitivity to these dataset characteristics.

The results of this investigation closely align with the findings of Caruana and Niculescu-Mizil's empirical analysis. Among the three supervised learning algorithms evaluated, SVM consistently had the highest performance in both this study and Caruana and Niculescu-Mizil's work. However, there were noteworthy deviations in the performances of KNN and Decision Tree compared to the original study. Specifically, Decision Tree significantly outperformed KNN on the more complex datasets, such as Wine and Breast Cancer, despite falling short of KNN's performance in Caruana and Niculescu-Mizil's empirical analysis.

5 Conclusion

This study compared the performance of SVM with RBF kernel, KNN, and Decision Tree across three datasets (Iris, Wine, and Breast Cancer Wisconsin). The results demonstrated that SVM consistently achieved the highest accuracy, with the Decision Tree performing nearly as well as SVM in most cases. KNN, while effective on simpler datasets like Iris, underperformed on complex datasets. The findings could potentially align more closely with Caruana and Niculescu-Mizil's work through improved strategies for addressing overfitting.

References

[1] Caruana, R. & Niculescu-Mizil, A. (2006) An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning (ICML '06). Association for Computing Machinery, New York, NY, USA, 161–168. https://doi.org/10.1145/1143844.1143865

[2] Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences. https://archive.ics.uci.edu/ml

[3] Fisher, R. (1936). Iris [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C56C76.

[4] Aeberhard, S. & Forina, M. (1992). Wine [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5PC7J.

[5] Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1993). Breast Cancer Wisconsin (Diagnostic) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B.