Performance of Supervised Learning Algorithms

Mohamed Abdilahi mabdilah@ucsd.edu

Link to Repository:
https://github.com/mabdilahCSE/ML-Two-Class-Classification-Project/tree/main

Abstract:

This report provides an in-depth comparison of three different machine learning classifiers--KNN, Decision Trees and Linear SVM. We assess their performance through the use of three diverse datasets: Cleveland Heart Disease, Census Income and Car Evaluation. Basically our goal was to empirically test these algorithms by constructing divided into two categories for each dataset. This project also emphasizes the need to choose appropriate machine learning models according to one's dataset features, results that may be of some assistance in future applications for use across many different fields. This comparative analysis helps increase understanding of how different algorithms perform under varying circumstances. Though there is still much room for such study, the results are invaluable to practitioners working within machine learning and data science fields.

Introduction:

This report delves into the comparative analysis of three prominent machine learning classifiers—K—Nearest Neighbors (KNN), Linear SVM, and Decision Trees—across three distinct datasets: Cleveland Heart Disease, Census Income & Car Evaluation. The intent of this study is to see how effective these classifiers are in different situations, giving new directions for data science and machine learning. In contrast with the KNN method, which is simple and effective in classification tasks, Linear SVM excels at handling high-dimensional data. For this reason, Decision Trees provide an intuitive and hierarchical approach to classification. Moreover, they can be used for the interpretation of results in cases where people want to know which factors are particularly important as causes or reasons leading up to a result.

The datasets selected cover a broad spectrum: medical diagnosis (Cleveland Heart Disease), socioeconomic factors (Census Income), and product quality assessment (Car

Evaluation). As the datasets emphasize different problems and hold their own peculiarities, they are suited to comprehensively testing the flexibility and adaptability of classifiers. This study is based on a review of the pertinent literature, in particular Caruana's landmark paper using empirical comparisons to analyze supervised learning algorithms. Caruana's research is particularly important for this report because it compares many different learning algorithms along a lot of different performance dimensions. Leveraging these lessons, the report therefore seeks to give a modern assessment of how classifiers get their classification intuition. This broad scope not only reinforces existing knowledge, it also examines new dimensions in applying such algorithms.

Method:

Three different datasets were used:

1. Heart Disease Dataset: A medical dataset about predicting whether or not there is heart disease according to various clinical and physiological factors.
2. Car Evaluation Dataset: This data set evaluates cars according to such factors as their buying price and maintenance cost, or safety.
3. Census Income Dataset: Concentrates on prediction of whether an individual's income is more than $50K/year based upon a person from the census data, which records age, education and occupation.

Classifiers:

1. K-Nearest Neighbors (KNN): KNN is a flexible classifier that was chosen for its simplicity, and proven results on different types of data.
2. Linear SVM: In high-dimensional spaces Linear SVM is particularly efficient. Its robustness, moreover, makes it suitable for a variety of situations in practice.
3. Decision Trees: This classifier offers a clear pathway to decision making, and so it is suitable for datasets with many different attributes.

Experimental Setup:

1. Preprocessing: Preprocessing was applied to each dataset, including treatment of missing values and normalization. Categorical data were recorded into numerical form using the label encoding method.
2. Parameter Settings: The parameters for each classifier were optimized with

GridSearchCV. KNN tests several 'K's, Linear SVM adjusts the regularization parameter C and Decision Trees max_depth.

3. Cross-Validation Approach: To make the results more general, a k-fold cross validation was performed to divide data into training and prediction sets

Experiment:

Performance Metrics:

Accuracy was used as the primary metric to evaluate model performance. Attaining high accuracy is widespread in studies of classification problems. It only counts the number of correctly predicted instances.

Results:

1. Heart Disease Dataset: KNN and Decision Trees both performed about as well, with Linear SVM following a little behind.
2. Car Evaluation Dataset: The former outperformed the latter by a significant margin. RTrees and KNN turned in outstanding performances, both clearly surpassing Linear SVM.
3. Census Income Dataset: Once again hiding the fact about its name, Decision Trees led all by a large margin. In second place was KNN followed way back in third (sadly) by Linear SVM.

After a comparative analysis, it appears that Decision Trees have generally had the best performance across datasets. KNN is second in most cases. The linear SVM, though effective, was invariably the least accurate.

Discussion:

According to the results, Decision Trees and KNN are more flexible for changing data types and complexities. One reason why decision trees work is that they are capable of processing non-linear relationships. KNN's remarkable record specifically betokens its flexibility in handling both categorical and continuous data.

This poorer performance of a linear SVM is most likely due to its flatness, which does not allow it as good an ability to reflect the complexity of larger datasets

Conclusion:

To sum up, through an observation of the performance distinction by KNN,Decision Trees and Linear SVM classifiers on three datasets (Heart Disease, Car Evaluation and Census Income), we are able to discover numerous interesting findings about these algorithms. The most robust and versatile proved to be Decision Trees, which not only bested others on many occasions in a variety of data situations but also increased accuracy through cross-validation. KNN tracked close behind, with dependable and flexible features. The linear SVM was effective, but relatively slow. This suggests that its power may be limited by the fact that it sometimes cannot handle complex datasets as efficiently as the other algorithms.

The main aim of subsequent research is to include more data and use a greater variety of datasets in the formulation stage, with real-life examples included. In addition, different classifiers and more datasets should be used if this study were to be expanded, which could potentially reaffirm current results. Moreover, it might turn up other clues to further research and understanding. In addition, it is possible to experiment with advanced preprocessing techniques and ideas for feature engineering in order to further improve the efficacy of classifiers. In conclusion, by deepening our comprehension of machine learning algorithms and how they work in real life, this study helps lead the way for further comprehensive research.

References:

Caruana, R., & Niculescu-Mizil, A. (2006). An Empirical Comparison of Supervised Learning Algorithms.

UCI Machine Learning Repository. Heart Disease Data Set.

UCI Machine Learning Repository. Car Evaluation Data Set

UCI Machine Learning Repository. Census Income Data Set