Statistical Comparison of Classifiers over Multiple Datasets

Giovanni Mantovani Andrea Palmieri

March 15, 2024

Abstract

This report presents a comprehensive statistical comparison of four classifiers—Support Vector Machine (SVM), Decision Tree, Random Forest, and K-Nearest Neighbors (K-NN)—across four distinct datasets. Our objective is to assess their performance and determine if any statistically significant differences exist between them. Employing a suite of tests tailored for multi-classifier and multi-dataset analyses, we evaluated classifier performance across diverse data scenarios. Despite variations in accuracies, the Friedman and Iman-Davenport tests revealed no statistically significant differences among the classifiers. The critical difference diagram further illustrated individual performance differences, highlighting the absence of a universally superior classifier.

1 Introduction

This report details a statistical comparison of four classifiers across a set of four datasets. To evaluate the performance and determine whether there are statistically significant differences between them, we implemented a series of tests for multi-classifier and multi-dataset comparisons.

2 Methodology

The analysis involved the following classifiers with their respective hyperparameters:

- Support Vector Machine (SVM) with linear kernel: Kernel scale of 1
- Support Vector Machine (SVM) with RBF kernel: Kernel scale of 0.1
- K-Nearest Neighbors (K-NN): Number of neighbors set to 10, using Euclidean distance
- Decision Tree (TREE): Gini diversity index (gdi) as split criterion, maximum number of splits set to 15

These classifiers were iteratively trained and tested across four different datasets. Stratified sampling was employed to divide each dataset into training and testing sets, ensuring that both sets are representative of the overall data, with a consistent class distribution.

2.0.1 Data Loading

The datasets are stored in separate MATLAB files. We load them iteratively. Once each dataset is loaded, we perform stratified sampling to divide it into training and testing sets. This process is carried out dataset by dataset, ensuring that for each one, the sampling results in training and testing sets that are representative of the original data, thereby maintaining the class distribution. After the data is split into training and testing sets, we proceed to train and evaluate various classifiers. We perform training on the training set and testing on the test set, and vice versa, to ensure the robustness of our evaluation.

2.1 Normality Test

To determine the appropriateness of parametric statistical methods such as ANOVA, we first assessed the normality of the data distribution in each dataset.

2.1.1 Theoretical Background

The normality test, specifically the Lilliefors test, is utilized to ascertain if data follow a normal distribution. This test is a modification of the Kolmogorov-Smirnov test, designed to be more suitable for situations where the sample's distribution mean and variance are not known. The null hypothesis of this test is that the data are normally distributed. Rejection of this hypothesis occurs when the p-value is less than the chosen significance level, typically 0.05, indicating non-normal distribution.

2.2 Ranking Classifier Accuracies

Ranking the accuracies of classifiers across datasets is a critical step in our analysis to understand their relative performance. We rank the classifiers based on their accuracy scores for each dataset, assigning the highest accuracy the rank of 1. In the event of tied accuracies, we distribute the average of the ranks.

2.3 Friedman Test for Classifier Comparison

To statistically compare the classifiers' performances, we employed the Friedman test.

2.3.1 Theoretical Background

The Friedman test is a non-parametric test used for comparing multiple observations across different groups. It's particularly useful when the data violates the normality assumption. The test ranks each observation in each group and compares these ranks across all groups to determine if there are significant differences.

The Friedman test calculates a chi-square statistic (χ_F^2) using the following formula:

$$\chi_F^2 = \frac{12N}{k(k+1)} \left(\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4} \right)$$

where:

- N is the number of datasets.
- k is the number of classifiers or treatments being compared.
- R_j is the sum of ranks for the j^{th} classifier across all datasets.

The test statistic follows a chi-square distribution with $k-1$ degrees of freedom. If the calculated χ_F^2 is greater than the critical value from the chi-square distribution table, given a significance level (usually 0.05), we reject the null hypothesis that all classifiers perform equally well.

2.4 Iman-Davenport Variation

2.4.1 Theoretical Background

The Iman-Davenport variation is a refined non-parametric statistical method tailored for detecting performance differences between groups in classifier comparisons. Unlike the Friedman test, it utilizes its own statistic, which is calculated by adjusting for the number of datasets and classifiers involved. This method enhances sensitivity, particularly in the context of machine learning analyses. The Iman-Davenport statistic is calculated as follows:

$$F_{ID} = \frac{(n-1) \cdot \chi_F^2}{n(k-1) - \chi_F^2} \tag{1}$$

where n is the number of datasets, k is the number of classifiers, and χ_F^2 is the Friedman Chi-square statistic. The resulting F_{ID} follows an F-distribution with $k-1$ and $(k-1)(n-1)$ degrees of freedom.

2.5 Critical Difference Calculation

Despite the Friedman test not revealing significant differences, we proceeded to calculate the Critical Difference (CD) to explore the extent of any disparities that might exist between the classifiers.

2.5.1 Theoretical Background

The CD for Nemenyi's post-hoc test quantifies the minimum difference in rank that is considered statistically significant. It's computed as:

$$CD = q_\alpha \times \sqrt{\frac{k \times (k+1)}{6 \times n}}$$

Where q_α is the critical value for the given alpha level, k is the number of classifiers, and n is the number of datasets. For alpha equal to 0.05 and 4 classifiers, $q_\alpha = 2.569$.

3 Results

3.1 Accuracies on Datasets

The following table summarizes the accuracies obtained by each classifier on the four datasets:

Dataset	SVM	Decision Tree	Random Forest	K-NN
Dataset 1	1.000	0.998	1.000	0.993
Dataset 2	0.883	0.808	0.872	0.836
Dataset 3	0.667	0.914	0.919	0.904
Dataset 4	0.562	0.957	0.944	0.959

Table 1: Accuracy of classifiers on different datasets.

3.2 Rankings of Classifiers

The performance rankings of the classifiers across the datasets are as follows:

Dataset	SVM	Decision Tree	Random Forest	K-NN
Dataset 1	1.5	3.0	1.5	4.0
Dataset 2	1.0	4.0	2.0	3.0
Dataset 3	4.0	2.0	1.0	3.0
Dataset 4	4.0	2.0	3.0	1.0

Table 2: Ranking of classifiers on different datasets.

3.3 Average and Final Rankings

The average and final rankings of the classifiers across all datasets are presented below. These rankings provide a comparative measure of each classifier's performance over the datasets.

Classifier	Average Ranking	Final Ranking
SVM	2.625	2
Decision Tree	2.750	3.5
Random Forest	1.875	1
K-NN	2.750	3.5

Table 3: Average ranking of classifiers across all datasets.

3.4 Normality Test Results

Using the Lilliefors test automated by a MATLAB function, we determined that not all datasets conformed to a normal distribution, prompting us to use non-parametric methods like the Friedman test for valid classifier performance analysis.

3.5 Friedman Test Results

The Friedman test assesses the null hypothesis that there are no significant differences in classifier performance.

Source	SS	df	MS	Chi-sq	Pvalue
Columns	2.1250	3	0.7083	1.3077	0.7273
Error	17.3750	9	1.9306	[ ]	[ ]
Total	19.5000	15	[ ]	[ ]	[ ]

Table 4: Results of the Friedman test for classifier comparison.

The Friedman test results indicate a Chi-square statistic of 1.3077 with a p-value of 0.7273. Given the high p-value, we do not reject the null hypothesis, suggesting that there are no statistically significant differences in the performances of the classifiers across the four datasets under consideration. This outcome suggests that the variations in classifier accuracies observed across different datasets may not be substantial enough to distinguish one classifier as consistently superior to the others.

3.6 Iman-Davenport Test Results

The Iman-Davenport test, with an $F_{ID} = 0.3669$ and a p-value of 0.7788, did not reveal significant differences among the classifier performances across multiple datasets, thereby confirming the results of the classical Friedman test. This consistency underlines that statistically significant variations in classifier efficiency are absent.

3.7 Critical Difference

Despite the non-significant results from the Friedman and Iman-Davenport tests, the critical difference diagram (see Figure 1) provides a visual examination of performance variations. In this diagram each algorithm is a bar centered in its average ranking with the width of the Critical Difference value, computed at 2.3452. While no algorithm consistently outperforms others, individual differences in performance still exist.

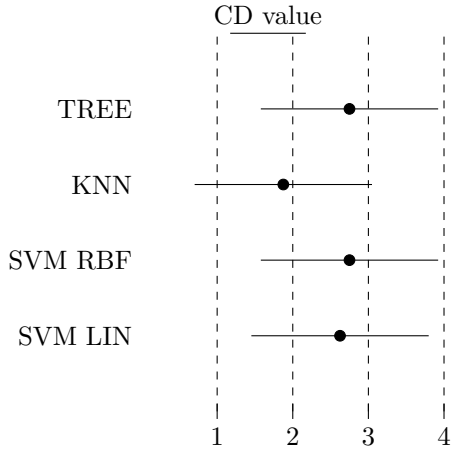

Figure 1: Evaluation and Comparison

4 Conclusion

In this study, we evaluated the performance of four different classifiers: SVM (with both linear and RBF kernels), K-NN, and Decision Tree across four datasets. Despite the individual performance variations observed, the statistical analysis through the Friedman and Iman-Davenport tests indicated no significant differences among the classifiers' overall performance. Our findings underscore the complexity of machine learning tasks and the importance of choosing the right classifier based on specific dataset characteristics

rather than presumed universal superiority. The Critical Difference diagram, although not showing significant statistical differences, visually highlight that no single classifier consistently outperformed the others. In summary, this study emphasizes the importance of choosing the right classifier for specific datasets in machine learning. It highlights that there's no one-size-fits-all solution and that the best approach depends on the unique characteristics of the data.