Waveform Dataset Project
Practical Application of Introduction to Machine Learning Techniques

Meryem Ben yahia [1]

Abstract

In this paper, I conducted a series of practical experiments on the Waveform Dataset (Breiman & Stone, 1988) to apply and test the theoretical concepts covered in the Introduction to Machine Learning course. The focus was on addressing the limitations of the k-nearest neighbor (kNN) algorithm to overcome its limitations.

1. Exploring the Waveform Dataset

The Waveform Dataset is a synthetic dataset that contains three distinct classes of waves, each characterized by 21 attributes. Its data is balanced as all of its classes are close in terms of ratio.

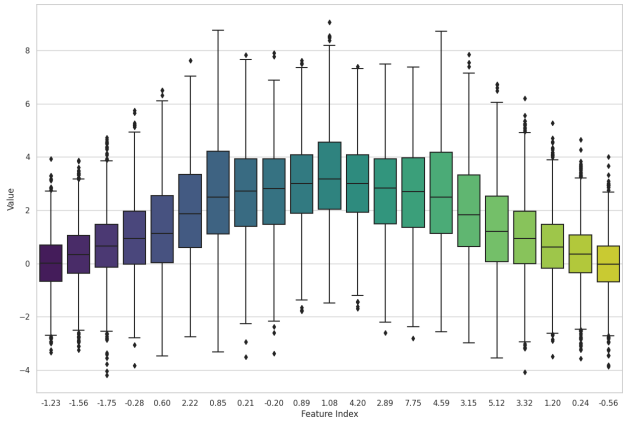

Figure 1. Box-plot showcasing the outliers per feature.

The box plot shows outliers in several features. Thus, the performance of the kNN can be impacted by those outliers since they can distort the distance calculations. The box heights also show varied degrees of spread, with the greatest variability within the middle 50% of the data. Thus, the performance of the kNN can be impacted by those outliers and it is also sensitive to this variability since they can distort the euclidean distance calculations.

To conclude, when using the k-Nearest Neighbors (kNN) algorithm on the Waveform dataset, a few important factors need to be taken into account. The complexity of the dataset—which includes mixed correlations, size variations, and outliers—means that preprocessing is required for kNN to perform well.

2. Experiments

2.1. Tuning the best K of a kNN classifier by Cross-validation

Accuracy increases with k up to a point, stabilizing thereafter. Thus, the performance of the kNN can be impacted by those outliers since they can distort the distance calculations. k=54 matches the Optimal Bayes Accuracy of 86%, indicating it as the optimal choice. Initially, lower k values are noise-sensitive, while larger ones average over neighbors.

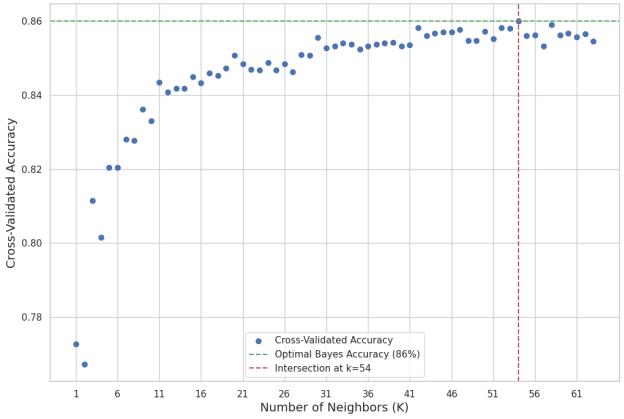

Figure 2. Accuracy vs. K for kNN.

*Equal contribution [1]Jean Monnet University, Saint-Étienne, France. Correspondence to: Meryem Ben yahia <meryem.ben.yahia@etu.univ-st-etienne.fr>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

2.2. Reducing the complexity of the kNN classifier

In this experiment, I reduced the complexity of waveform dataset, first by reducing the size while preserving its key features, and adding the CNN rule for a better generalization of the k-Nearest Neighbors (k-NN) model. The objective was for the model focus on essential patterns in the data, to hopefully achieve accuracy on unseen data, such as the test set.

Table 1. Classification accuracy by applying data reduction algorithms to the training data and then comparing the accuracy of a 1-NN classifier on the 1000 test waves before and after the reduction.

EXPERIMENT	ACCURACY	BETTER?
1NN (BEFORE REDUCTION)	77.20	
REDUCED TRAINING SET	89.60	√
REDUCED TRAINING AND CNN	89.70	√

Using these methods, the accuracy gain was evident in the reported test accuracy rates (89.60% and 89.70%), which also surpass the Optimal Bayes Classification baseline of 86%.

2.3. Sped-up calculation of the 1NN

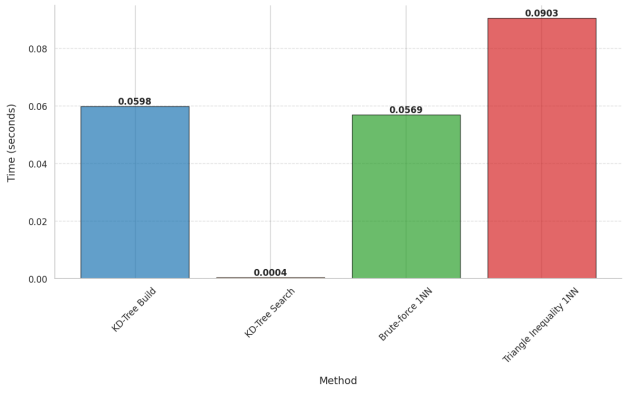

Figure 3. Performance Comparison of 1NN Search Approaches.

From this experiment and comparing 1NN Bruteforce, KD-Trees, and Triangle inequality acceleration, we conclude:

- Although KD-Tree build time adds an initial cost that is quickly offset by its fast search speed, its lower efficiency and rebuilding costs make it unsuitable for high-dimensional or dynamic datasets. KD-Tree search runs significantly faster than brute-force, which makes it perform well on for moderately sized static datasets.
- Brute-force 1NN, despite being the slowest and least scalable, remains practical for small or dynamic datasets where the overhead of more complex structures is not justifiable.
- While the triangle inequality optimization is an improvement over brute-force search, it is still not as effective as the KD-Tree approach, especially when there are a lot of points that are equally far from the query point.

For the waveform dataset, the KD-Tree search offers the best trade-off between build time and query speed, suitable for static datasets, while brute-force remains a fallback for its simplicity.

2.4. Generating artificial imbalancy, and tuning K using the F-measure

2.4.1. ARTIFICIAL IMBALANCY

In this experiment, Class 1 is lowered by 50% and class 2 by 75% in this updated training dataset. The accuracy on the test set decreases from a 77.2% on a balanced dataset to a less desirable 74.5% on the artificially imbalanced dataset. The results obtained from inducing an artificial imbalance supported a key-characteristic of the performance of the k-Nearest Neighbors (kNN) classifier. The decline in accuracy observed in the case of the imbalanced training set can be primarily attributed to kNN's inherent bias towards the majority class during the training process.

2.4.2. TUNING K USING THE F-MEASURE

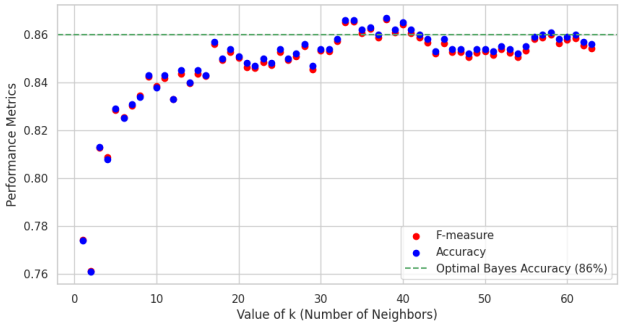

Figure 4. Performance Comparison: Accuracy vs F-measure for k.

Figure 4 shows that after an initial increase, accuracy and F-measure for the k-Nearest Neighbors classifier improve and stabilize. It indicates that within the studied k range, there is a good compromise between overfitting and generalization. Through the creation of artificial minority class samples, SMOTE helped to mitigate class imbalance and contribute to this balanced performance.

Acknowledgements

I wish to express my heartfelt appreciation to Dr. Sebban for his exceptional guidance and instruction throughout this semester.

I would also like to express my appreciation to the MLDM 2023–2025 class for supplying an exciting place to learn. My learning experience has been tremendously enhanced and my perspectives have been substantially expanded by the conversations, teamwork, and idea-sharing within our class.

Thank you.

References

Breiman, L. and Stone, C. Waveform Database Generator (Version 1). UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C5CS3C.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.