Analyzing the Performance of the k-Nearest Neighbor Classifier Using Metrics Based on Neighborhood Sequences

Oleksandra Polinski[a], **Attila Fazekas**[b]

[a]Faculty of Informatics, University of Debrecen, Ferenc Rakoczi II Transcarpathian Hungarian University
palinszky.alexandra@inf.unideb.hu

[b]Faculty of Informatics, University of Debrecen,
attila.fazekas@inf.unideb.hu

Abstract

The k-Nearest Neighbor (kNN) classifier is a simple yet effective supervised learning algorithm for classification tasks. Being a non-parametric method, its performance is highly influenced by the choice of distance metrics. Classical distance metrics do not always provide the desired level of classification accuracy, making it worthwhile to explore alternative metrics that may improve classifier performance on specific datasets [1, 3].

The aim of the study is to investigate the applicability of neighborhood sequence-based distance metrics [2] in the k-Nearest Neighbor (kNN) algorithm, and subsequently to compare the performance of the resulting classifiers with those based on classical distance metrics. In the evaluation, commonly used performance measures such as accuracy, recall, precision, and F1-score are considered. Balanced, publicly available datasets with varying sample sizes and dimensionality are employed to assess classifier performance. It is important to highlight that dataset imbalance can also significantly affect the performance of the kNN classifier; however, this is not the primary focus of the present study. As a preprocessing step, all datasets are rescaled to ensure that differences in feature scales do not influence the results [4].

The performance of the kNN classifier is examined on several real-world datasets using ten different distance metrics, with performance measures evaluated in each case. Based on the results, neighborhood sequence-based distance metrics provide better performance than classical distance metrics such as Euclidean, Manhattan, and Chebyshev on certain datasets.

Based on the results of testing the kNN algorithm on the datasets, it can be stated that neighborhood sequence-based distance metrics perform well, although they do not achieve the best performance on the datasets examined. Consequently, neighborhood sequence-based distance metrics, such as the octagonal distance metric, may provide a viable alternative for improving the performance of kNN classifiers on real-world datasets. In the future, it may be worthwhile to investigate the application of these distance metrics with other classification algorithms. Furthermore, extending the study to datasets with larger sample sizes and higher dimensionality would be advisable.

References

[1] H. A. ALFEILAT, A. O. AHMAD HASSANAT: *Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review*, Big Data 7.4 (2019), pp. 221–248, DOI: 10.1089/big.2018.0175.

[2] A. FAZEKAS, A. HAJDU, L. HAJDU: *Metrical neighborhood sequences in Z*, Pattern Recognition Letters 26.13 (2005), pp. 2022–2032, DOI: 10.1016/j.patrec.2005.02.006.

[3] L. HU, M. HUANG, S. KE, et al.: *The distance function effect on k-nearest neighbor classification for medical datasets*, SpringerPlus 5.1304 (2016), DOI: 10.1186/s40064-016-2941-7.

[4] M. YUSRAN, K. SADIK, A. M. SOLEH, C. SUHAENI: *Effect of Feature Normalization and Distance Metrics on K-Nearest Neighbors Performance for Diabetes Disease Classification*, Journal of Mathematics, Computations, and Statistics 8.2 (2025), pp. 341–354, DOI: 10.35580/jmathcos.v8i2.8012.