

Comparative Analysis of KNN and KNN_CLAS: Statistical Performance and Likelihood Space Equivalence

Eduardo Henrique Basilio de Carvalho
Departamento de Engenharia Eletrônica
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
eduardohbc@ufmg.br

Abstract—This report presents a comparative analysis of the standard K-Nearest Neighbors (KNN) classifier and a variant, KNN_CLAS, which utilizes expert support points derived from Gabriel graphs. The comparison is conducted on three publicly available datasets: Spect Heart, Ionosphere, and Haberman’s Survival. We evaluate their statistical performance using metrics like accuracy, precision, recall, and F1-score, with significance testing (Wilcoxon signed-rank and paired t-tests) under cross-validation. Furthermore, we analyze the geometry of their likelihood spaces (q_0, q_1 scores) using metrics such as centroid distance and Bhattacharyya distance. The experiments consider K values of {1, 3, 11} and Gaussian kernel bandwidth H values of {0.01, 0.1, 1.0}. Results indicate that KNN_CLAS is not universally superior but shows comparable or significantly better performance than KNN under specific hyperparameter settings (notably $H=1.0$ for some datasets), suggesting its potential for computational efficiency via expert point selection. The spatial analysis reveals differences in likelihood space geometries, which do not always directly correlate with cross-validated performance, highlighting the complex interplay between model architecture, hyperparameter settings, and generalization.

Index Terms—pattern recognition, large margin classifiers, Gabriel graph, KNN classifier, statistical analysis, likelihood space, KNN_CLAS.

I. INTRODUCTION

The K-Nearest Neighbors (KNN) algorithm is a fundamental non-parametric method used for classification and regression. Its simplicity and intuitive nature make it a popular choice. However, its performance can be sensitive to the choice of K, distance metric, and the presence of noisy or irrelevant data points. Variants of KNN aim to address these limitations. One such approach involves informed selection of training points, akin to support vector concepts in large-margin classifiers.

The Classifier by Support Edges (CLAS) methodology, often leveraging Gabriel graphs, aims to identify critical “expert” points that define the decision boundary [1]–[4]. KNN_CLAS, the focus of this study, is a KNN variant that uses such expert points, derived from Gabriel graph edges connecting samples of different classes (support edges). This study investigates the equivalence between a standard KNN (using a Gaussian kernel with Mahalanobis distance) and KNN_CLAS. The comparison

is twofold: 1) statistical analysis of classification performance metrics, and 2) analysis of the spatial characteristics of their respective likelihood (q_0, q_1) probability spaces. The goal is to understand if KNN_CLAS provides comparable or improved performance, potentially with computational advantages due to its reliance on a reduced set of “expert” points, and how their decision-making mechanisms differ in likelihood space. The underlying research draws from concepts of large margin classifiers and geometric data representation [5]–[7].

II. METHODOLOGY

A. Classifiers

Two classifiers were compared:

- **KNN**: A K-Nearest Neighbors classifier using a Gaussian kernel with Mahalanobis distance. The covariance matrix for the Mahalanobis distance is computed from the training data.
- **KNN_CLAS**: This variant first fits a standard KNN model to derive the covariance structure. It then identifies “expert” points by constructing a Gabriel graph on the training data and selecting points that form “support edges” (Gabriel edges connecting points of different classes). Predictions are then made using a KNN approach considering only these expert points, also with a Gaussian kernel and the previously computed Mahalanobis distance.

Both classifiers utilize hyperparameters K (number of neighbors) and H (bandwidth for the Gaussian kernel). The values tested were $K \in \{1, 3, 11\}$ and $H \in \{0.01, 0.1, 1.0\}$.

B. Datasets

The analysis was performed on three datasets obtained from the UCI Machine Learning Repository [8]:

- 1) Spect Heart
- 2) Ionosphere
- 3) Haberman’s Survival

C. Preprocessing

A standard preprocessing pipeline was applied to each training set within the cross-validation folds:

- 1) **VarianceThreshold**: Removal of features with variance below 1×10^{-3} .
- 2) **CorrelationFilter**: Removal of features with an absolute correlation coefficient greater than 0.9 with a preceding feature.
- 3) **StandardScaler**: Standardization of features by removing the mean and scaling to unit variance.

D. Statistical Performance Analysis

A K-Fold cross-validation strategy was employed, with $N_{splits} = \min(30, N_{samples})$, where $N_{samples}$ is the number of samples in the dataset. For each fold, models were trained on the training partition and evaluated on the test partition. Performance was measured using:

- Accuracy
- Precision (with zero_division=0)
- Recall (with zero_division=0)
- F1-score (with zero_division=0)

The scores from each fold were collected, and then, for each (Dataset, K, H) combination, the performance of KNN was compared against KNN_CLAS using:

- **Wilcoxon signed-rank test**: A non-parametric test for paired samples.
- **Paired t-test**: A parametric test for paired samples. Cohen's d was calculated as an effect size for significant t-test results, where the difference was taken as (KNN scores - KNN_CLAS scores).

A p-value less than 0.05 was considered statistically significant.

E. Spatial Likelihood Analysis

For each (Dataset, K, H) combination, the classifiers were trained on the entire preprocessed dataset. Then, likelihood scores q_0 (sum of kernels for class -1 neighbors) and q_1 (sum of kernels for class +1 neighbors) were computed for all training samples. Based on these (q_0, q_1) points, the following spatial metrics were calculated to characterize the geometry of the likelihood space:

- **Centroid Distance**: Euclidean distance between the centroids of (q_0, q_1) points belonging to class 0 and class 1.
- **Mean Distance Opposite**: Mean Euclidean distance between (q_0, q_1) points from opposite classes.
- **Mean Distance Same**: Mean Euclidean distance between (q_0, q_1) points from the same class.
- **Bhattacharyya Distance**: A measure of similarity between the two distributions of (q_0, q_1) points for class 0 and class 1.
- Variance of q_0 and q_1 scores for each class.

It is important to note that this spatial analysis was conducted on the entire training dataset, not within a cross-validation scheme.

III. RESULTS

The analysis generated extensive results for statistical performance and spatial likelihood metrics across all datasets and hyperparameter combinations. Due to space constraints, we summarize key findings and present illustrative examples.

A. Statistical Performance

Table I and Table II summarize notable statistical comparisons for the Spect Heart and Haberman Survival datasets. P-values indicate the significance of the difference between KNN and KNN_CLAS (KNN - KNN_CLAS scores).

Generally, for $H = 0.01$ and $H = 0.1$, KNN and KNN_CLAS showed few statistically significant differences in performance across most metrics and datasets (p-values often > 0.05). However, for $H = 1.0$, more distinct differences emerged.

TABLE I
SELECTED STATISTICAL TEST RESULTS FOR SPECT HEART (P-VALUES).
NEGATIVE COHEN'S D INDICATES KNN_CLAS PERFORMED BETTER.

Metric	K	H	Wilcoxon p	t-test p	Cohen's d	Sig. Diff.
Accuracy	1	1.0	0.0014	0.0007	-0.695	Yes (CLAS > KNN)
Recall	1	1.0	0.0014	0.0002	-0.789	Yes (CLAS > KNN)
F1-score	1	1.0	0.0017	0.0007	-0.689	Yes (CLAS > KNN)
Accuracy	3	1.0	0.0045	0.0064	-0.536	Yes (CLAS > KNN)
Recall	3	1.0	0.0006	0.0001	-0.813	Yes (CLAS > KNN)
F1-score	3	1.0	0.0027	0.0040	-0.570	Yes (CLAS > KNN)
Accuracy	11	1.0	0.0025	0.0006	-0.705	Yes (CLAS > KNN)
Recall	11	1.0	<0.0001	<0.0001	-1.061	Yes (CLAS > KNN)
F1-score	11	1.0	0.0002	0.0001	-0.796	Yes (CLAS > KNN)

For the **Spect Heart** dataset (Table I), when $H = 1.0$, KNN_CLAS consistently and significantly outperformed KNN across accuracy, recall, and F1-score for all tested K values. For $H = 0.01$ and $H = 0.1$, no significant differences were observed (p-values ≈ 1.0).

For the **Ionosphere** dataset with $H = 1.0$:

- For K=1, no significant differences were found (e.g., accuracy Wilcoxon p ≈ 0.317).
- For K=3, KNN showed marginally better accuracy (Wilcoxon p ≈ 0.040 , t-test p ≈ 0.056 , Cohen's d ≈ 0.363).
- For K=11, no significant strong differences were observed (e.g. accuracy Wilcoxon p ≈ 0.197).

For $H = 0.01$ and $H = 0.1$ on Ionosphere, no significant differences were found.

TABLE II
SELECTED STATISTICAL TEST RESULTS FOR HABERMAN SURVIVAL
(P-VALUES). POSITIVE COHEN'S D INDICATES KNN PERFORMED BETTER.

Metric	K	H	Wilcoxon p	t-test p	Cohen's d	Sig. Diff.
Accuracy	1	0.01	0.0031	0.0007	0.690	Yes (KNN > CLAS)
Recall	1	0.01	0.0656	0.0998	-0.310	No (Trend CLAS > KNN)
Accuracy	1	0.1	0.0335	0.0314	0.413	Yes (KNN > CLAS)
Accuracy	1	1.0	0.0335	0.0314	0.413	Yes (KNN > CLAS)
Accuracy	3	0.01	0.0031	0.0008	0.683	Yes (KNN > CLAS)
Accuracy	11	0.01	0.0030	0.0008	0.683	Yes (KNN > CLAS)

For the **Haberman Survival** dataset (Table II), KNN often performed significantly better or equivalently. For instance, with $H = 0.01$, KNN showed significantly better accuracy for all K values. For $H = 0.1$ and $H = 1.0$ with $K=1$, KNN also had better accuracy.

B. Spatial Likelihood Analysis

The spatial analysis provided insights into the geometry of the (q_0, q_1) likelihood space on the full training set. Table III shows an example for Spect Heart ($K=1, H=1.0$).

TABLE III
SPATIAL LIKELIHOOD METRICS FOR SPECT HEART ($K=1, H=1.0$)

Metric	KNN	KNN_CLAS
Centroid Distance	1.29×10^{-7}	5.40×10^{-8}
Bhattacharyya Dist.	2.07×10^{-6}	3.64×10^{-7}
Mean Dist. Opposite	1.29×10^{-7}	6.16×10^{-8}
Mean Dist. Same	0.0	4.23×10^{-8}
Var q_0 (avg)	3.50×10^{-46}	5.52×10^{-16}
Var q_1 (avg)	8.76×10^{-47}	1.02×10^{-15}

For Spect Heart with $K=1$ and $H=1.0$ (Table III), where KNN_CLAS demonstrated superior statistical performance (Table I), KNN exhibited larger centroid and Bhattacharyya distances in the likelihood space computed on the full training set. This is a counter-intuitive finding if one assumes larger separation in likelihood space directly implies better generalization.

For very small H values (e.g., $H=0.01$), some spatial metrics for both classifiers, particularly on Spect Heart and Ionosphere, showed extremely large magnitudes (e.g., 10^{24} to 10^{37}). This may suggest numerical sensitivity or extreme localization of the Gaussian kernel with very small bandwidths, potentially making these specific spatial metrics less interpretable in those regimes. For example, for Spect Heart ($K=1, H=0.01$), KNN had a centroid distance of $\approx 1.29 \times 10^{37}$, while KNN_CLAS had $\approx 5.40 \times 10^{36}$.

Across datasets and other K, H combinations, the relationship between the magnitudes of these spatial metrics and cross-validated statistical performance was not always direct or consistent. For instance, on the Ionosphere dataset ($K=3, H=1.0$), where KNN had slightly better accuracy, its Bhattacharyya distance was $\approx 1.72 \times 10^{-8}$ compared to $\approx 1.29 \times 10^{-8}$ for KNN_CLAS.

IV. DISCUSSION

The comparative analysis of KNN and KNN_CLAS reveals a nuanced relationship regarding their performance and likelihood space characteristics.

Statistical Performance Equivalence: The two methods are not universally equivalent in performance. For smaller Gaussian kernel bandwidths ($H = 0.01, H = 0.1$), KNN and KNN_CLAS generally exhibit statistically similar performance across the tested metrics and datasets. This suggests that in regimes where the kernel has a very localized influence, the expert selection mechanism of KNN_CLAS does not

significantly alter performance compared to standard KNN. In such cases, KNN_CLAS might be preferable due to potential computational savings from using fewer points for prediction.

However, with a larger bandwidth ($H = 1.0$), significant performance differences emerged. On the **Spect Heart** dataset, KNN_CLAS demonstrated clear superiority over KNN. This implies that for this dataset and bandwidth, the "expert" points selected by KNN_CLAS are more effective for generalization than relying on all training points as KNN does. Conversely, on the **Haberman Survival** dataset, KNN often outperformed KNN_CLAS, particularly for $H = 0.01$ and also for $H = 0.1, H = 1.0$ at $K=1$. For the **Ionosphere** dataset, results were mixed, with KNN sometimes showing marginal advantages or no significant difference.

These results suggest that the effectiveness of KNN_CLAS's expert selection strategy is dataset-dependent and highly influenced by the kernel bandwidth H . A larger H allows for broader influence of neighboring points; in this scenario, the quality of selected "expert" points by KNN_CLAS becomes crucial and can lead to either improved or degraded performance compared to KNN.

Spatial Likelihood Analysis Insights: The analysis of the (q_0, q_1) likelihood space provides a view into how the classifiers represent class separability based on the training data. A key observation is that superior spatial separation metrics (like centroid distance or Bhattacharyya distance) on the *entire training set* do not always translate to superior *cross-validated statistical performance*. For example, in the Spect Heart case ($K=1, H=1.0$), KNN showed larger spatial separation metrics but was outperformed by KNN_CLAS in cross-validation.

This discrepancy highlights an important distinction: the spatial likelihood analysis was performed on the full training set after model fitting, while statistical performance was assessed via cross-validation on unseen test folds. The global geometry of the likelihood space on training data might not fully capture the model's generalization capabilities. KNN_CLAS, by focusing on "expert" points near the presumed decision boundary, might achieve better generalization even if its global likelihood space on training data appears less separated by some metrics.

The extremely large values for some spatial metrics at $H = 0.01$ are a concern, possibly indicating that the Gaussian kernel normalization combined with very small bandwidth and the dataset's covariance structure leads to values that are numerically unstable or dominate the calculations, making comparisons difficult in this specific regime.

Overall Equivalence and Implications: KNN and KNN_CLAS are not equivalent under all conditions. KNN_CLAS appears to be a promising alternative, particularly when computational efficiency is a concern, as it often performs comparably to KNN (especially for smaller H values). Its potential to outperform KNN (as seen with Spect Heart for $H=1.0$) suggests that its mechanism of selecting support points can be beneficial. However, it can also underperform, indicating that the expert selection is not a universally optimal

strategy.

The KNN_CLAS model's behavior of adjusting K if the number of expert points is less than the specified K is a factor that could influence comparisons, especially if very few experts are found.

Future work could explore adaptive methods for choosing H or more robust spatial metrics, particularly for scenarios with very small bandwidths. Investigating the characteristics of the "expert" points selected by KNN_CLAS across different datasets and H values could also provide deeper insights.

In conclusion, KNN_CLAS offers a valuable modification to the standard KNN by incorporating concepts from Gabriel graphs and support-based learning. Its performance is competitive and sometimes superior, but its advantages are context-dependent, relying on the dataset characteristics and hyperparameter choices, particularly the kernel bandwidth H . The spatial likelihood analysis, while informative, should be interpreted cautiously as a direct predictor of generalization performance.

ACKNOWLEDGMENT

This work utilizes datasets from the UCI Machine Learning Repository [8].

REFERENCES

- [1] L. A. Torres, "Classificador por arestas de suporte (clas): métodos de aprendizado baseados em grafos de gabriel," *Manuscript (PhD thesis)*, 2016.
- [2] A. C. Souza, C. Leite Castro, J. A. Garcia, L. C. B. Torres, L. J. Acevedo Jaimes, and B. R. A. Jaimes, "Improving the efficiency of gabriel graph-based classifiers for hardware-optimized implementations," in *2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA)*, 2019, pp. 1–5.
- [3] J. Arias-Garcia, L. C. B. Torres, M. A. Castrillon-Franco, G. Osorio, and G. Castellanos-Dominguez, "Enhancing performance of gabriel graph-based classifiers by a hardware co-processor for embedded system applications," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 2, pp. 1265–1275, 2021.
- [4] J. Arias-Garcia, L. C. B. Torres, D. Campo-Muñoz, M. A. Castrillón-Franco, and G. Castellanos-Dominguez, "Improved design for hardware implementation of graph-based large margin classifiers for embedded edge computing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 1, pp. 1075–1089, 2024.
- [5] L. C. B. Torres, C. L. Castro, and A. P. Braga, "A parameterless mixture model for large margin classification," in *2015 International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–8.
- [6] L. C. B. Torres, C. L. Castro, F. Coelho, and A. P. Braga, "Large margin gaussian mixture classifier with a gabriel graph geometric representation of data set structure," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 3, pp. 1000–1012, 2021.
- [7] L. C. B. Torres, C. L. Castro, F. Coelho, F. Sill Torres, and A. P. Braga, "Distance-based large margin classifier suitable for integrated circuit implementation," Universidade Federal de Minas Gerais, Manuscript, Tech. Rep., 2015.
- [8] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: <http://archive.ics.uci.edu/ml>