Comparative Analysis of KNN and KNN_CLAS: 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 focuses on the geometry of their likelihood spaces (q_0, q_1 scores) across several publicly available datasets, including Spect Heart, Ionosphere, and Haberman's Survival. We analyze metrics such as centroid distance, Bhattacharyya distance, mean distance between points from opposite and same classes, and variance of q_0/q_1 scores to understand the equivalence or divergence in their likelihood representations. The experiments consider K values of {1, 3, 11} and Gaussian kernel bandwidth H values of {0.01, 0.1, 1.0}. The analysis aims to shed light on how the underlying mechanisms of these classifiers, particularly the expert point selection in KNN_CLAS, influence the structure of their likelihood spaces, independent of traditional performance metrics. This offers insights into their decision-making processes.**

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

I. Introduction

The K-Nearest Neighbors (KNN) algorithm is a fundamental non-parametric method used for classification. Its simplicity and intuitive nature make it a popular choice. Variants of KNN aim to address its limitations, such as sensitivity to noisy data or the choice of K. One such approach involves an informed selection of training points, drawing inspiration from support vector concepts in large-margin classifiers. The Classifier by Support Edges (CLAS) methodology, often leveraging Gabriel graphs, seeks to identify critical "expert" points that are influential in defining the decision boundary. KNN_CLAS, a focus of this study, is a KNN variant that employs such expert points, which are derived from Gabriel graph edges connecting samples of different classes (referred to as support edges).

This study investigates the characteristics of the likelihood spaces generated by a standard KNN (using a Gaussian kernel with Mahalanobis distance) and KNN_CLAS. The core of this analysis lies in comparing the spatial properties of their respective likelihood (q_0, q_1) probability spaces. The q_0 score represents the sum of Gaussian kernel influences from neighbors belonging to class -1, while the q_1 score represents the sum for class +1 neighbors. The distribution of these (q_0, q_1) pairs for training samples forms a 2D space that reflects how the classifier separates classes. The goal is to understand if, and under what conditions, KNN and KNN_CLAS produce equivalent or differing geometries in this likelihood space, providing insights into their decision-making mechanisms rather than their classification performance. This research draws upon concepts related to large margin classifiers and the geometric representation of data structures.

II. Methodology

A. Classifiers

Two classifiers were compared:

- **KNN**: A K-Nearest Neighbors classifier that utilizes a Gaussian kernel with Mahalanobis distance. The covariance matrix required for the Mahalanobis distance is computed from the entire training dataset.
- **KNN_CLAS**: This variant initially fits a standard KNN model to establish the covariance structure from all training data. Subsequently, it identifies "expert" points by constructing a Gabriel graph on the training data. Points that form "support edges" — Gabriel edges connecting data points of different classes — are selected as these experts. Predictions are then made using a KNN approach that considers only these selected expert points, employing the same Gaussian kernel and the previously computed Mahalanobis distance. If the number of identified expert points is less than the specified K, KNN_CLAS adjusts K to be the number of available expert points for that prediction instance.

Both classifiers are parameterized by K (the number of neighbors) and H (the bandwidth for the Gaussian kernel). The experiments were conducted with $K \in \{1, 3, 11\}$ and $H \in \{0.01, 0.1, 1.0\}$.

B. Datasets

The analysis was performed on multiple datasets, with a focus on three obtained from the UCI Machine Learning Repository [1] for detailed discussion:

1) Spect Heart
2) Ionosphere
3) Haberman's Survival

Other datasets included in the full analysis were Breast Cancer, Pima Diabetes, Digits Binary, and Sonar.

C. Preprocessing

A consistent preprocessing pipeline was applied to each dataset before training the classifiers. This pipeline consisted of the following steps, executed in order:

1) **VarianceThreshold**: Features with a variance below 1×10^{-3} were removed. This step helps eliminate features that are nearly constant across samples.
2) **CorrelationFilter**: Features exhibiting an absolute correlation coefficient greater than 0.9 with any preceding feature were removed. This reduces multicollinearity.
3) **StandardScaler**: The remaining features were standardized by removing the mean and scaling to unit variance. This ensures that all features contribute more equally to distance computations.

D. Spatial Likelihood Analysis

For each dataset and each combination of hyperparameters (K, H), the classifiers were trained on the entire preprocessed dataset. Following training, likelihood scores q_0 (sum of kernel influences from neighbors of class -1) and q_1 (sum of kernel influences from neighbors of class +1) were computed for all training samples using the respective model's 'likelihood_score' method. These (q_0, q_1) points form a two-dimensional likelihood space.

To characterize the geometry of this space for each classifier, the following spatial metrics were calculated:

- **Centroid Distance**: The Euclidean distance between the centroid of (q_0, q_1) points belonging to class -1 and the centroid of points belonging to class +1.
- **Mean Distance Opposite Classes**: The average Euclidean distance between (q_0, q_1) points that originate from samples of opposite classes.
- **Mean Distance Same Class**: The average Euclidean distance between (q_0, q_1) points that originate from samples of the same class.
- **Bhattacharyya Distance**: A measure of the similarity (or divergence) between the two distributions of (q_0, q_1) points, one for each class. This metric considers both the mean and covariance of the two distributions.
- **Variance of q_0 scores (Var q_0)**: The average variance of the q_0 scores, computed across points belonging to class -1 and class +1 separately, then averaged if both are present.
- **Variance of q_1 scores (Var q_1)**: Similar to Var q_0, but for the q_1 scores.

This spatial analysis was conducted on the full training set for each configuration to understand the intrinsic data representation learned by the models in their likelihood space.

III. Results

The spatial likelihood analysis yielded a comprehensive set of metrics for each dataset, classifier, and hyperparameter combination. Due to space limitations, we present selected results that highlight key trends and differences in the likelihood space geometries.

Table I shows the spatial likelihood metrics for the Spect Heart dataset with K=1 and H=1.0.

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

Metric	KNN	KNN_CLAS
Centroid Distance	1.288×10^{-7}	5.396×10^{-8}
Bhattacharyya Distance	2.075×10^{-6}	3.639×10^{-7}
Mean Dist. Opposite	1.288×10^{-7}	6.158×10^{-8}
Mean Dist. Same	0.0	4.233×10^{-8}
Var q_0 (avg)	3.503×10^{-46}	5.517×10^{-16}
Var q_1 (avg)	8.758×10^{-47}	1.017×10^{-15}

Table II presents metrics for the Ionosphere dataset with K=3 and H=1.0.

TABLE II
SPATIAL LIKELIHOOD METRICS FOR IONOSPHERE (K=3, H=1.0)

Metric	KNN	KNN_CLAS
Centroid Distance	1.174×10^{-8}	1.017×10^{-8}
Bhattacharyya Distance	1.722×10^{-8}	1.294×10^{-8}
Mean Dist. Opposite	1.196×10^{-8}	1.052×10^{-8}
Mean Dist. Same	3.375×10^{-9}	3.342×10^{-9}
Var q_0 (avg)	1.646×10^{-18}	2.302×10^{-18}
Var q_1 (avg)	6.313×10^{-18}	6.017×10^{-18}

Table III shows metrics for Haberman's Survival dataset with K=11 and H=0.1.

TABLE III
SPATIAL LIKELIHOOD METRICS FOR HABERMAN'S SURVIVAL (K=11, H=0.1)

Metric	KNN	KNN_CLAS
Centroid Distance	3.330	2.123
Bhattacharyya Distance	2.916	2.337
Mean Dist. Opposite	3.575	2.519
Mean Dist. Same	1.443	1.258
Var q_0 (avg)	1.282	0.904
Var q_1 (avg)	0.279	0.257

When H is very small (e.g., H=0.01), several spatial metrics, particularly for KNN, exhibited extremely large magnitudes for datasets like Spect Heart and Ionosphere, sometimes reaching orders of 10^{20} to 10^{30} or higher for centroid distance and Bhattacharyya distance. For example, for Spect Heart with K=1, H=0.01, KNN's centroid distance was $\approx 1.29 \times 10^{37}$ and its Bhattacharyya distance was $\approx 1.57 \times 10^{29}$. In contrast, for the same configuration, KNN_CLAS had a centroid distance of $\approx 5.40 \times 10^{36}$ and a Bhattacharyya distance of ≈ 94.30 (Table IV). This suggests that for KNN with very small

H, the likelihood scores can become extremely separated or scaled, potentially due to the Gaussian kernel becoming highly localized and sensitive to the data's covariance structure. KNN_CLAS, by using a potentially different set of (expert) neighbors, sometimes mitigates this effect on certain metrics like the Bhattacharyya distance, though its centroid distances can also be very large.

TABLE IV
SPATIAL LIKELIHOOD METRICS FOR SPECT HEART (K=1, H=0.01) ILLUSTRATING LARGE MAGNITUDES

Metric	KNN	KNN_CLAS
Centroid Distance	1.288×10^{37}	5.396×10^{36}
Bhattacharyya Distance	1.566×10^{29}	9.430×10^{1}
Mean Dist. Opposite	1.288×10^{37}	6.158×10^{36}
Mean Dist. Same	0.0	4.233×10^{36}
Var q_0 (avg)	6.969×10^{41}	5.517×10^{72}
Var q_1 (avg)	6.969×10^{41}	1.017×10^{73}

Across different datasets and hyperparameter settings, the relationship between the two classifiers in terms of these spatial metrics varied. No single classifier consistently produced "larger" or "smaller" values across all metrics or all conditions. This indicates that the geometry of the likelihood space is sensitive to the choice of K, H, the dataset itself, and the specific mechanism of the classifier (all points vs. expert points).

IV. DISCUSSION

The analysis of spatial likelihood metrics provides insights into how KNN and KNN_CLAS structure their decision spaces based on the q_0 and q_1 scores. The primary goal was to assess the equivalence of these spaces under various conditions.

Equivalence and Divergence in Likelihood Spaces: The likelihood spaces of KNN and KNN_CLAS are not universally equivalent. For larger H values (e.g., H=1.0), the metrics in Tables I and II show that KNN and KNN_CLAS can produce quantitatively different geometries. For Spect Heart (K=1, H=1.0), KNN exhibits larger Centroid Distance and Bhattacharyya Distance than KNN_CLAS. This suggests that, on the full training set, KNN's likelihood space shows greater separation between the class centroids and overall distributions in this specific configuration. However, for Ionosphere (K=3, H=1.0), while KNN still has slightly larger centroid and Bhattacharyya distances, the values are closer in magnitude to those of KNN_CLAS.

When the bandwidth H is moderate (e.g., H=0.1, as in Table III for Haberman's Survival), both classifiers yield spatial metrics of similar orders of magnitude, though still distinct values. This suggests that their likelihood spaces, while not identical, might share some structural similarities in terms of class separability as captured by these metrics. The most striking divergences appear at very small H values ($H = 0.01$) [**?**], [**?**]. As seen in Table IV, KNN can produce extremely large values for centroid distance and Bhattacharyya distance. KNN_CLAS also shows very large centroid distances and variances for q_0/q_1 under these conditions, but its Bhattacharyya distance for Spect Heart ($K = 1$, $H = 0.01$) was notably smaller than KNN's by many orders of magnitude. This indicates that with a highly localized kernel (small H), the selection of specific neighbors (all for KNN vs. experts for KNN_CLAS) can lead to vastly different characteristics in the computed likelihood distributions. The extreme values for KNN might suggest that its q_0, q_1 points for the two classes are pushed to extreme regions of the likelihood space, or that the distributions become ill-conditioned for metrics like Bhattacharyya distance, possibly due to near-zero kernel activations for most training points outside a tiny radius. The `batched_gaussian_kernels` function uses $\sum_{\text{scaled}}^{-1} = \frac{\sum^{-1}}{h^2}$, so a small h dramatically increases the values in the exponent's quadratic form, potentially leading to extreme kernel values if not properly normalized or if distances are small.

Influence of Classifier Mechanics: The use of "expert" points by KNN_CLAS is a key differentiator. These points are selected based on Gabriel graph support edges, intending to capture critical instances near the decision boundary. When H is large, the Gaussian kernel has a broad influence. In this scenario, the choice of neighbors (all vs. experts) can significantly alter the summed kernel influences (q_0, q_1). If expert points provide a more "focused" representation of class boundaries, this could lead to a differently structured likelihood space compared to using all points. When H is very small, the kernel is highly localized. KNN, considering all points, might find very few effective neighbors contributing to q_0, q_1 scores, and these could be highly sensitive to noise or outliers, leading to extreme metric values. KNN_CLAS, by pre-selecting expert points, might operate on a more stable, albeit smaller, set of influential points, which could explain why its Bhattacharyya distance (which depends on covariance estimates) was less extreme in some H=0.01 cases (e.g., Spect Heart).

The observation that KNN_CLAS adjusts its K value if the number of experts is less than the initially specified K is also relevant. If K is effectively reduced for KNN_CLAS in many instances, this inherently means it is using information from fewer neighbors to compute q_0, q_1 scores compared to KNN using the full K, which would naturally lead to different likelihood space geometries.

Interpreting Spatial Metrics: It is crucial to reiterate that these spatial metrics describe the geometry of the likelihood space on the training data. A larger centroid or Bhattacharyya distance might suggest better class separation *in that specific 2D likelihood representation of the training set*. However, this does not directly predict generalization performance on unseen data. A model might achieve good separation on the training set's likelihood space but overfit, or conversely, a more compact likelihood space might still generalize well if the decision boundary learned is robust. The current analysis

focuses only on the geometric properties themselves as a basis for comparing the operational characteristics of the classifiers.

The extremely large values for metrics at H=0.01 warrant caution. They may indicate numerical instability or that the assumptions underlying metrics like Bhattacharyya distance (e.g., well-behaved covariance matrices) are challenged when kernels become delta-like.

Conclusion on Equivalence: KNN and KNN_CLAS do not generally produce equivalent likelihood spaces. The degree of similarity or divergence is highly dependent on the dataset characteristics and, crucially, the hyperparameter H. For very small H, their likelihood spaces can be dramatically different, especially for metrics sensitive to distribution shape like Bhattacharyya distance. For moderate to larger H, the differences might be more nuanced but still present. The expert selection mechanism of KNN_CLAS fundamentally alters the set of points contributing to the likelihood scores, leading to distinct geometric configurations in the (q_0, q_1) space compared to standard KNN. Further investigation could explore the topological properties of these spaces or how the density of points from each class is distributed, to gain even deeper insights into their equivalence.

ACKNOWLEDGMENT

This work utilizes datasets from the UCI Machine Learning Repository [1]. The conceptual basis for KNN_CLAS and related Gabriel graph techniques draws from works such as [2]–[8].

REFERENCES

[1] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[2] L. A. Torres, "Classificador por arestas de suporte (clas): métodos de aprendizado baseados em grafos de gabriel," *Manuscript (PhD thesis)*, 2016.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.