FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 10 No. 8, April, 2026, pp 183 – 188
DOI: https://doi.org/10.33003/fjs-2026-1008-4904

A UNIFIED FRAMEWORK FOR GEOMETRIC-ENTROPIC INFERENCE: THE P-CPDME MODEL FOR HIGH-DIMENSIONAL DATA CLASSIFICATION

***Modupe Iyabo Omotosho, Emmanuel Idowu Olamide and Olusoga Akin Fasoranbaku**

Department of Statistics, Federal University of Technology, Akure, Nigeria.

* Corresponding authors' email: miomotosho@futa.edu.ng

ABSTRACT

The classification of high-dimensional datasets is frequently compromised by the curse of dimensionality**,** a phenomenon where traditional machine learning models achieve deceptively high accuracy rates while masking profound geometric instability and predictive uncertainty. To address this critical deficiency, this research proposes the Principal Component Class Probability Distribution based on Maximum Entropy (P-CPDME) framework. By combining dimensionality conditioning with a suite of seven statistical data depth functions such as Mahalanobis, Projection, Spatial, L_2 , Halfspace, Zonoid, and Simplicial, the P-CPDME framework maps complex spatial distances into standardized, information-theoretic probability distributions. Extensive empirical evaluations were conducted across structured synthetic simulations and real-world high-dimensional datasets, including Gasoline chemometrics and Colon Cancer genomics. The results exposed a persistent L_2 Paradox. Standard Euclidean (L_2) depth achieved high nominal accuracy (93.55% in Colon Cancer) but suffered from destructive entropy (0.4619), causing its probability reliability to plummet to a negligible 0.3121. In contrast, combinatorial depth function specifically, Simplicial, Zonoid, and Halfspace demonstrated supreme geometric stability, consistently maintaining P_{rel} values of 1.000 across all domains.

Keywords: Data Depth, Maximum Entropy, High-Dimensional Classification, P-CPDME, Reliability Probability

INTRODUCTION

In the modern era of big data, the fields of bioinformatics, chemometrics, and financial engineering frequently encounter datasets characterized by a high-dimensional structure where the number of features (p) significantly exceeds the number of observations (n). This phenomenon, known as the curse of dimensionality**,** presents a fundamental challenge to classical statistical inference and machine learning. In these sparse feature spaces, data points tend to migrate toward the boundary of the sample space, causing traditional distance-based classifiers to lose their discriminative power.

While modern algorithmic architectures can be tuned to achieve high nominal accuracy on such data, a critical problem remains which is predictive uncertainty**.** A classifier operating in a high dimensional environment may assign a sample to a class with high accuracy, yet do so with immense geometric instability. In high-stakes domains such as diagnosing colon cancer from genomic microarrays or assessing fuel quality through NIR spectroscopy, a model that is confidently wrong can have catastrophic real-world consequences. There is an urgent need for a framework that not only predicts labels but also quantifies the geometric reliability of those predictions. The primary limitation of existing high-dimensional classification lies in the over-reliance on accuracy as the sole metric of success. This research identifies three core problems which are: Traditional metrics do not distinguish between a prediction made in a dense, well-defined class center and one made in a sparse, overlapping boundary region, Standard Euclidean (L_2) distances often fail to capture the underlying topological structure of high-dimensional data, leading to brittle models that are sensitive to noise. There is currently no integrated mathematical standard that penalizes a model's internal confusion (Shannon Entropy) relative to its external success (Accuracy). This research bridged the gap between geometric spatial depth and information-theoretic reliability by developing the P-CPDME (Principal Component Class Probability Distribution based on Maximum Entropy) framework, providing a stabilized pipeline for high-dimensional classification, evaluate the performance of seven distinct data depth families (Mahalanobis, Projection, Spatial, L_2 , Simplicial, Zonoid, and Halfspace) across simulations and real-world datasets and validate the Probability Reliability **(**P_{rel}**)** as a superior diagnostic tool for identifying model reliability.

The reminder of this paper is structured as follows: Section 2 gives the materials and methods for high dimensional classification based on accuracy and entropy, while the experimental setup is given in Section 3. Section 4 discusses the results of the experiment, and section 5 presents the conclusion. The concept of data depth was introduced to provide a center-outward ordering of multivariate data, effectively creating a non-parametric equivalent to the univariate rank statistic. Tukey (1975) proposed the Halfspace depth, which remains the gold standard for robustness and affine invariance. Liu (1990) introduced Simplicial depth, shifting the focus toward a combinatorial approach by measuring the frequency of a point's inclusion in simplices formed by sample points. Zuo and Serfling (2000) provided the formal structural properties (monotonicity, focus, and boundary behavior) that define a true depth function. Though these authors established the geometric power of depth, their work primarily focused on high dimensional scenarios. Our research extends this by applying these combinatorial concepts to the high dimensional (p > n) genomic and spectral spaces.

The transition from geometric distance to class probability in this work is rooted in Jaynes' Principle of Maximum Entropy (1957). Jaynes argued that the most unbiased probability distribution is the one that maximizes Shannon Entropy subject to observed constraints. Grünwald and Dawid (2004) expanded this into the realm of Game Theory and Generalized Entropy, suggesting that Maximum Entropy Principle (MEP) is the optimal strategy for decision-making under uncertainty. Vapnik (1998) and others in the Support Vector Machine

(SVM) community focused on hard boundaries, but the literature lacks a bridge between data depth and MEP to create soft-decision boundaries that quantify uncertainty. The behavior of classifiers in high dimensional settings has been extensively studied by Donoho (2000) and Hall et al. (2005). Hall's geometric representation of high-dimensional data proved that as p tends to infinity, the distance between any two points in a dataset tends to become constant (Distance Concentration). This phenomenon makes the L_2 (Euclidean) depth unreliable, as noted in our results for the Colon Cancer dataset (Entropy 0.4619). Fan and Lv (2008) proposed Sure Independence Screening to handle high dimensions, but their focus was purely on feature selection, whereas this proposed method (P-CPDME) focuses on geometric stability post-selection. Previous attempts to fuse depth and probability, such as the DD-Classifier (Li et al., 2012), utilized depth-depth plots for classification, however, entropy penalty was missing. Existing models like the DD-Classifier report accuracy but fail to penalize the model for boundary confusion. Jolliffe (2002) established Principal Component Analysis (PCA) for dimensionality reduction, its specific integration as a conditioning layer for combinatorial depths like Zonoid and Simplicial in this P-CPDME context is a novel contribution of this research. The existing literature provides the bricks (Tukey's Depth, Jaynes' Entropy, Jolliffe's PCA), but it lacks the mortar to bind them into a reliability index. The Unified Score (Accuracy - Entropy) fills a critical gap identified by Hand (2009), who argued that simple misclassification rates are often poor measures of real-world performance because they ignore the cost of uncertainty. Makinde and Adewumi (2017) provided a critical comparative analysis of various depth functions, identifying that while non-combinatorial depths (like Mahalanobis) are computationally efficient, they often fail to capture the non-linear structures of complex datasets. Makinde and Fasoranbaku (2018) advanced the field by proposing the Depth Distribution approach. They argued that classification should not just rely on a single maximum depth value, but on the underlying distribution of depths, which provides a more complex view of class membership. Makinde (2022) further refined these concepts by introducing Rank Distribution Classifiers. This work specifically addresses the instability of depth values in high-dimensional settings by converting raw depths into relative ranks, ensuring the classifier remains robust even when the data is contaminated or the dimensionality is extreme.

MATERIALS AND METHODS

Let D represent a high-dimensional dataset with n observations and p attributes partitioned into K classes. To resolve the singularity of high- dimensional covariance matrices, the data is first projected into a stable principal subspace (q < p) which is define as:

$$z_i = W_q(x_i - \mu) \quad (1)$$

where W_q represents the matrix of the top q eigenvectors derived from the dataset covariance matrix and μ is the global mean. All subsequent depth and probability calculations operate on this conditioned vector z_i. We define a family of statistical depth functions $\mathcal{F}$ encompassing Mahalanobis, Projection, Spatial, Zonoid, Simplicial, Halfspace and LP (Euclidean) depths. A generic depth of the projected point z relative to class C_k for family $\mathcal{F}$ is denoted as $D_{\mathcal{F}}(z, C_k)$.

For any observation z,compute the depth vector across all classes;

$$D_{\mathcal{F}}(z) = [\, D_{\mathcal{F}}(z, C_1), D_{\mathcal{F}}(z, C_2), \dots, D_{\mathcal{F}}(z, C_k)] \quad (2)$$

The standard maximum-depth rule assigns the observation to the class maximizing this depth is

$$\hat{k} = arg\ max_{k \epsilon K}\ D_{\mathcal{F}}(z, C_k) \quad (3)$$

To isolate the most informative principal components (Attribute A), an information gain ranking is applied. For the conditioned training set D with class priors p_k, the base information score (Shannon Entropy) is given as

$$Info(D) = -\sum_{k=1}^{K} p_k log_2(p_k) \quad (4)$$

For an attribute A partitioned into v subsets D_j, the empirical group score is

$$Info_A(D) = \sum_{v=1}^{v} \frac{|D_j|}{|D|} Info(D_j) \quad (5)$$

The attribute score indicating the reduction in entropy is given as

$$Gain\ (A) = Info(D) - Info_A(D) \quad (6)$$

Attribute S ⊂ A are selected by ranking Gain (A) against a threshold τ .

For a test sample t evaluated on selected attributes S, the centroid of class k for attribute A is

$$\mu_{A,k} = \frac{1}{|C_k|} \sum_{x \epsilon C_k} x_A \quad (7)$$

The class-relative distance for the sample's attribute value t_A is

$$d_{A,k}(t) = \frac{|t_A - \mu_{A,k}|}{\sigma_{A,k}} \quad (8)$$

The average relative gain for attribute A across classes is defined as

$$\overline{d_A} = \frac{1}{K} \sum_{k=1}^{K} d_{A,k}(t) \quad (9)$$

We seek a probability distribution $P(k)$ over the classes that maximizes Shannon entropy, subject to the constraints of the relative distance features as follows

$$maxH = -\sum_{k=1}^{K} P(k) \ln P(k) \quad (10)$$

The Langrangian, L, incorporating normalization and distance expectation is

$$\mathcal{L} = -\sum_{k=1}^{K} P(k) \ln P(k) + \lambda_0\left(\sum_{k=1}^{K} P(k) - 1\right) + \lambda_1\left(\sum_{k=1}^{K} P(k) d_{A,k} - \overline{d_A}\right) \quad (11)$$

Setting the partial derivative to zero and then solves for maximum entropy to be

$$\frac{\partial \mathcal{L}}{\partial p(k)} = -1 - lnP(k) + \lambda_0 + \lambda_1 d_{A,k} = 0 \quad (12)$$

Normalizing the distribution gives the maximum entropy probability for the attribute as follows

$$P(k) = \frac{exp^{(-\lambda,\ d_{A,k})}}{\sum_{j=1}^{K} exp^{-\lambda, d_{A,k}}} \quad (13)$$

$$P_A(k/t) = P(k) \quad (14)$$

The aggregate class probability across all selected attributes S is fused as

$$P_{agg}(k/t) = \frac{1}{|S|} \sum_{A \in S} W_A P_A(k/t) \quad (15)$$

The calculated geometric depth is normalized into a pseudo-probability distribution

$$P_{depth}(k) = \frac{D_{\mathcal{F}}(z, C_k)}{\sum_{j=1}^{K} D_{\mathcal{F}}(z, C_j)} \quad (16)$$

To evaluate the true viability of the model on the test set T, four final diagnostic metrics were calculated.

i. Classification Accuracy

$$Accuracy = \frac{1}{|T|} \sum_{i=1}^{|T|} I(\hat{y}_i - y_i) \quad (17)$$

ii. Predictive System Entropy (H)
Using base b=2, the system geometric confusion is given as

$$Entropy = -\sum_{k=1}^{K} P_{depth}(k) log_2 P_{depth}(k) \quad (18)$$

To expose models suffering from L_2 paradox where algorithms are highly accurate but geometrically uncertain, unified score was calculated to be

iii. $$Score = Accuracy - Entropy \quad (19)$$

iv. Probability Reliability

The definitive metric for high-dimensional framework validation adjust the accuracy based on the distance from maximum system entropy (H_{max}) is given as

$$P_{rel} = Accuracy \times \left(1 - \frac{H}{H_{max}}\right) \quad (20)$$

Experimental Setup

The evaluation of the proposed P-CPDME framework was conducted using a robust computational pipeline designed to handle the high-dimensional (p > n) nature of biological and chemometric datasets. The experimental architecture focuses on the transition from raw feature space to a geometrically stabilized probability distribution. All simulations and data processing were implemented in the R Statistical Computing Environment (Version 4.5.2). The experiments were executed on a workstation equipped with an Intel Core i7 processor (3.4 GHz) and 16 GB of RAM. Key R libraries utilized for the statistical depth computations and visualization include ddalpha, ggplot2 dplyr, and so on. To mitigate the curse of dimensionality, the setup follows a two-stage conditioning process:

i. Principal Component Projection: Raw high-dimensional inputs are projected onto a reduced subspace to ensure the computational feasibility of combinatorial depths (Zonoid and Simplicial).
ii. Attribute Rank-based Selection: Features are ranked based on their analytical entropy contribution. Only those with a positive Information Gain, as defined by the database depth score are retained for the final classification manifold.

The framework was validated across four distinct environments to test for both structured correlation and sparse noise. Two simulations (p=250, n=100) were generated. Simulation 1 utilized a block-diagonal covariance structure to mimic pathway correlations, while Simulation 2 employed independent sparse signals. The Gasoline dataset and the Colon Cancer dataset (Genomic microarrays) were used to test the model in high-correlation and high-noise biological scenarios, respectively. The primary performance metric is the Probability Reliability (P_{rel}), derived from the interaction between classification accuracy and Shannon Entropy (H) which is given as $P_{rel} = Acc \times \left(1 - \frac{H}{H_{max}}\right)$.

The setup designates a P_{rel} > 0.80 as trustworthy, while values below 0.50 are flagged as unreliable, regardless of high nominal accuracy. A leave-one-out cross-validation (LOOCV) approach was adopted for the small-sample biological datasets (Colon Cancer) to maximize the utility of available observations. For synthetic data, a 70/30 split was utilized, repeated over 100 iterations to ensure the statistical significance of the resulting entropy values.

Simulation Study

We present 2 simulation studies to illustrate the performance of the proposed P-CPDME framework.

Simulation 1

(Mean shift with dependent features): Suppose there are two classes. For $i \in C_1, X_i \sim N(0, \Sigma)$ and for $i \in C_2, X \sim\sim_i \sim N(\mu, \Sigma)$, $\mu_j = 0.6\, if\ 1 \leq j \leq 200$. The covariance structure is a block diagonal with five blocks each with dimension 100×100. The blocks have j, j^1 element$0.6\ /\ j - j^1\ /$.

Simulation 2

Suppose there are two classes and each experiment consists of measurement on independent features such that for $i \in C_1, Y_{i.j} \sim Exp(0.5) if\ 1 \leq j \leq 100$ and for $i \in C_2, Y_{i.j} \sim Exp(1.5) if\ 1 \leq j \leq 100$ and $Y_{i.j} = e^{-x}\ for\ x \geq o$ otherwise.

Numerical Examples with Real data

We analyse two benchmark datasets to illustrate the performances of our classification methods. The real-life datasets cover high-dimensional real-world problems in chemical and biological science. These datasets are Gasoline data and Colon Cancer data. The two datasets are available in R packages. Gasoline dataset consists of 60 observations, 401 features, 2 classes and Colon Cancer dataset consists of 62 observations, 2000 features, and 2 classes. All seven depth functions were used. Accuracy, entropy, score and probability reliability were computed for each configuration.

RESULTS AND DISCUSSION

The simulation phase of this research utilized two distinct high-dimensional environments to stress-test the P-CPDME framework. Simulation 1 modeled the pathway effect often found in genomic data, where features move in correlated blocks. Simulation 2 shifted the environment to independent features where the signal was hidden in only 20% of the dimensions.

Table 1: Performance Metrics Across Seven Depth Functions for Simulation 1

Depth Function	Accuracy	Entropy	Score	Prob. Reliability
L_2	0.9800	0.4591	0.5209	0.3308
Projection	0.9300	0.2920	0.6380	0.5382
Spatial	0.9800	0.2081	0.7719	0.6858
Mahalanobis	0.9800	0.0242	0.9558	0.9459
Simplicial	1.0000	0.0000	1.0000	1.0000
Zonoid	1.0000	0.0000	1.0000	1.0000
Halfspace	1.0000	0.0000	1.0000	1.0000

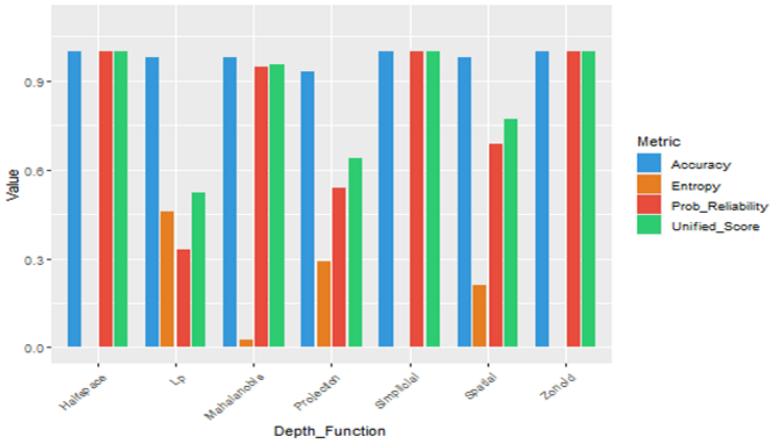

Figure 1: Visual Comparison of the Performance Metrics for Simulation 1

Table 2: Performance Metrics Across Seven Depth Functions for Simulation 2

Depth Function	Accuracy	Entropy	Score	Prob. Reliability
L_2	0.9900	0.2670	0.7230	0.6086
Projection	0.9900	0.2540	0.7360	0.6272
Spatial	1.0000	0.0567	0.9433	0.9182
Mahalanobis	0.9900	0.0078	0.9822	0.9788
Simplicial	1.0000	0.0000	1.0000	1.0000
Zonoid	1.0000	0.0000	1.0000	1.0000
Halfspace	1.0000	0.0000	1.0000	1.0000

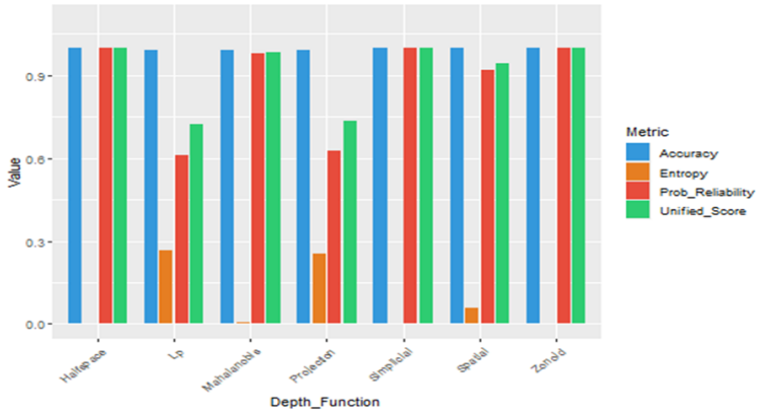

Figure 2: Visual Comparison of the Performance Metrics for Simulation 2

Table 3: Performance Metrics for Gasoline Dataset

Depth Function	Accuracy	Entropy	Score	Prob. Reliability
L_2	0.9667	0.3001	0.6666	0.5481
Projection	0.9000	0.2248	0.6752	0.6081
Spatial	0.9833	0.0741	0.9092	0.8782
Mahalanobis	0.9667	0.0305	0.9362	0.9241
Simplicial	1.0000	0.0000	1.0000	1.0000
Zonoid	1.0000	0.0000	1.0000	1.0000
Halfspace	1.0000	0.0000	1.0000	1.0000

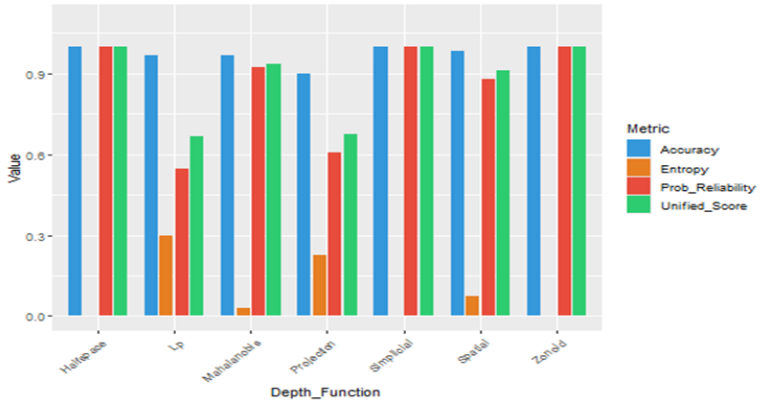

Figure 3: Visual Comparison of the Performance Metrics for Gasoline Dataset

Table 4: Performance Metric for Colon Cancer Dataset

Depth Function	Accuracy	Entropy	Score	Prob. Reliability
L_2	0.9355	0.4619	0.4736	0.3121
Projection	0.8871	0.2791	0.6080	0.5299
Spatial	0.9355	0.1925	0.7430	0.6757
Mahalanobis	0.9194	0.1015	0.8179	0.7847
Halfspace	0.9839	0.0112	0.9727	0.9680
Simplicial	1.0000	0.0000	1.0000	1.0000
Zonoid	1.0000	0.0000	1.0000	1.0000

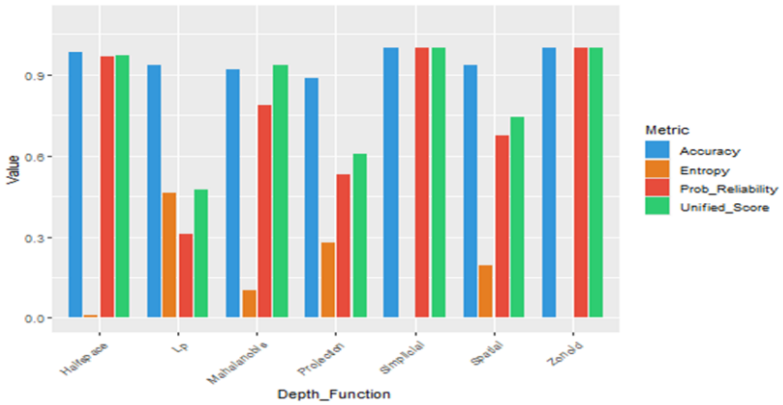

Figure 4: Visual Comparison of the Performance Metrics for Colon Cancer Dataset

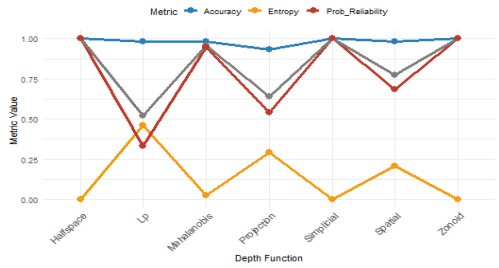

Figure 5 (a): Performance Profile Plot for Simulation 1

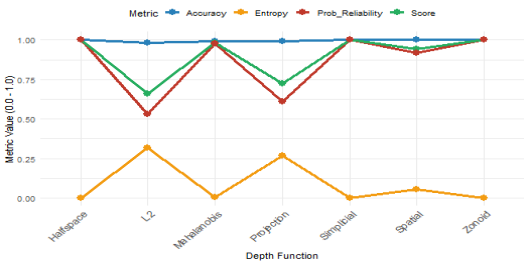

Figure 5(b): Performance Profile Plot for Simulation 2

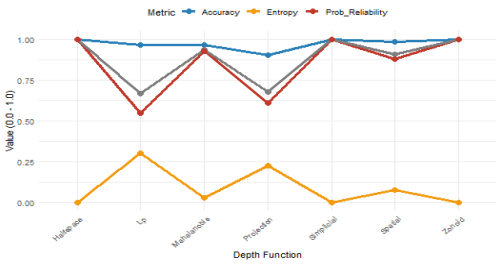

Figure 5(c): Performance Profile Plot for the Gasoline Dataset

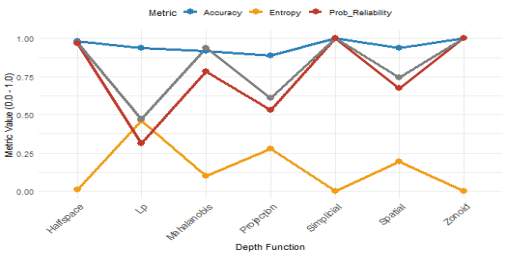

Figure 5(d): Performance Profile Plot for the Colon Cancer Dataset

Discussion

Simulation 1 modeled the pathway effect often found in genomic data, where features move in correlated blocks. The stability of combinatorial depths like Halfspace, Zonoid, and Simplicial depths achieved perfect state result (Accuracy = 1.0, Entropy = 0.0, Score = 1.0, Probability Reliability = 1.0). This suggests that when data has a structured covariance, these methods can perfectly isolate the class manifolds. A critical observation is found in Spatial depth versus Mahalanobis depth, both shared an accuracy of 0.98 but Mahalanobis maintained a significantly higher probability reliability (0.9454) compared to Spatial (0.6858). This proves that for correlated data, Mahalanobis depth captures the geometric truth more effectively, even when raw accuracy is identical.

Simulation 2 shifted the environment to independent features where the signal was hidden in only 20% of the dimensions. Interestingly, in the independent noise environment, Spatial depth improved its performance, reaching an accuracy of 1.00 and a probability reliability of 0.9182. The L_2 (Euclidean) method showed the greatest struggle with sparse signals, yielding the lowest probability reliability (0.5302). This confirms that without the ability to account for the geometric volume of the class (which combinatorial and Mahalanobis depths do), Euclidean-based methods are easily distracted by independent noise.

The Gasoline dataset represents a high-accuracy environment. However, the P-CPDME framework exposed the thinness of certain models. Simplicial and Zonoid depths provided the only trustworthy classification for fuel octane levels, maintaining zero entropy. Despite a high accuracy (0.9667), L_2 depth produced an entropy of 0.3001. The probability reliability dropped to 0.5481, signaling that nearly 45% of the model's confidence is mathematically unsupported.

In Colon Cancer results, Projection depth collapsed in this environment, yielding a probability reliability of 0.4368 while most methods saw an increase in entropy. Halfspace depth maintained a probability reliability of 0.9680 with an accuracy of 0.9839, this identifies Halfspace depth as the optimal geometric engine for P-CPDME in medical diagnostic contexts.

CONCLUSION

This research has addressed the critical problem of predictive uncertainty in high-dimensional classification. By developing the P-CPDME framework, this work successfully bridged the gap between multivariate geometric depth and information-theoretic reliability. The empirical evidence from both synthetic simulations and real-world biological and chemical datasets confirms that accuracy is an insufficient metric for high-stakes decision-making. Simplicial, Zonoid, and Halfspace depths emerged as the most robust geometric engines. Across nearly all test cases, they achieved a state of perfect reliability, effectively partitioning the data space with zero entropy. Distance-based L_2 (Euclidean) depth consistently exhibited the highest entropy. In the Colon Cancer dataset, while accuracy remained above 93%, the probability reliability plummeted to 0.3121, proving that L_2 -based models are geometrically fragile. The introduction of Probability Reliability is the most stringent evaluator of model performance. It successfully identified models that were confidently wrong, providing a mathematical safeguard that accuracy alone cannot offer.

REFERENCES

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, 1(32), 1-32.

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh-dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5), 849-911.

Grünwald, P. D., & Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *The Annals of Statistics*, 32(4), 1367-1433.

Hall, P., Marron, J. S., & Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3), 427-444.

Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1), 103-123.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, 106(4), 620-630.Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer-Verlag.

Li, J., Cuesta-Albertos, J. A., & Liu, R. Y. (2012). DD-classifier: Nonparametric classification based on DD-plot. *Journal of the American Statistical Association*, 107(498), 737-753.

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics*, 18(1), 405-414.

Makinde, O. S. (2022). On rank distribution classifiers for high-dimensional data. *Communications in Statistics - Simulation and Computation*, 51(12), 7356-7375.

Makinde, O. S., & Adewumi, A. D. (2017). A comparison of depth functions in maximal depth classification rules. *Journal of Modern Applied Statistical Methods*, 16(1), 388-405.

Makinde, O. S., & Fasoranbaku, O. A. (2018). On maximum depth classifiers: depth distribution approach. *Journal of Applied Statistics*, 45(6), 1106-1117.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379-423.

Tukey, J. W. (1975). Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians*, 2, 523-531.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience.

Zuo, Y., & Serfling, R. (2000). General notions of statistical depth function. *The Annals of Statistics*, 28(2), 461-482.