Active Learning Classification from a Signal Separation Perspective

Hrushikesh Mhaskar*†, Ryan O'Dowd† and Efstratios Tsoukanis†

Institute of Mathematical Sciences, Claremont Graduate University, Claremont, CA, 91711
Emails: Hrushikesh.Mhaskar@cgu.edu, ryan.o'dowd@cgu.edu, efstratios.tsoukanis@cgu.edu

***Abstract*—In machine learning, classification is usually seen as a function approximation problem, where the goal is to learn a function that maps input features to class labels. In this paper, we propose a novel clustering and classification framework inspired by the principles of signal separation. This approach enables efficient identification of class supports, even in the presence of overlapping distributions. We validate our method on real-world hyperspectral datasets Salinas and Indian Pines. The experimental results demonstrate that our method is competitive with the state of the art active learning algorithms by using a very small subset of data set as training points.**

I. Introduction

Classification is one of the oldest and most extensively studied problems in machine learning. Mathematically, the problem can be formulated as follows: we consider data of the form $\{(x, y)\}$, where $x \in \mathcal{D}$ for some domain $\mathcal{D}$ (e.g., Euclidean space or graph vertices), and y takes values in a finite set, conveniently encoded as $\{1, \ldots, K\}$ for some integer $K \geq 2$. In the pair (x, y), y is the **class label** of x, and the **classification function** is defined as $f(x) = y$.

Over the past 50 years, machine learning research has produced numerous algorithms for estimating the class label of any data point $x \in \mathcal{D}$, all of which approximate the function f. This perspective unifies the classification problem with the classical problem of function approximation.

Approximation theory focuses on methods for estimating real-valued functions defined on subsets of Euclidean space and analyzing the intrinsic (non-statistical) errors associated with such approximations. Typically, the target function f is assumed to be "smooth" on its domain. Although the classification function is inherently piecewise constant and discontinuous, this is not a theoretical obstacle when classes are well-separated. In such cases, extension theorems by Stein [1] guarantee the existence of infinitely differentiable extensions of piecewise constant functions to the entire Euclidean space, even preserving the magnitude of the derivatives.

In modern applications, classes often overlap, and even when they are disjoint, their boundaries may lack smoothness, posing challenges for classical function approximation techniques in classification. Additionally, extension theorems provide only existence results without offering constructive methods to obtain such extensions, particularly when class boundaries are unknown. We propose that classification can be effectively approached through an analogy to the problem of signal separation in phased array antennas.

In this problem, one wants to determine a linear combination $\mu = \sum_{k=1}^{K} a_k \delta_{\omega_k}$, (where δ_ω denotes the Dirac delta distribution supported at ω), given the Fourier coefficients $\hat{\mu}(\ell) = \sum_{k=1}^{K} a_k \exp(-i\ell\omega_k)$ for finitely many values of ℓ, say $|\ell| < n$ for some integer n. One way to solve this problem is to consider a smooth, even, low pass filter H supported on $[-1, 1]$, and construct

$$\sigma_n(\mu)(x) = \sum_{\ell \in \mathbb{Z}} H(|\ell|/n)\hat{\mu}(\ell)\exp(i\ell x).$$

Under certain conditions, we have shown in [2], [3], [4], [5] that $\sigma_n(\mu) \approx \mu$, so that the set $\{x : |\sigma_n(\mu)(x)| > \min_k |a_k|/2\}$ splits into exactly K clusters and the maxima of $|\sigma(\mu)(x)|$ in these

[*] Supported in part by ONR grant N00014-23-1-2394.
[†] Supported in party by ONR grant N00014-23-1-2790.

clusters occur very close to the points ω_k. This is because $\sigma_n(\mu)$ is a convolution of μ with the kernel $\Phi_n^T(y) = \sum_{\ell} H(|\ell|/n)\exp(iky)$, and this kernel is highly localized when H is smooth.

To connect with classification, assume $a_k \geq 0$ for all k with $\sum_k a_k = 1$. Suppose the data belongs to classes $\{\omega_1, \ldots, \omega_K\}$, where $x = \omega_k$ implies class label k. Here, μ represents the data distribution, and each class k is sampled from δ_{ω_k}.

Generalizing, each class k corresponds to a distribution μ_k, making the overall data distribution a convex combination:

$$\mu^* = \sum_{k=1}^{K} a_k \mu_k.$$

The technical barriers in this viewpoint are the following: (1) the supports of μ_k may be continuous rather than discrete, and (2) we observe random samples from μ^* instead of Fourier coefficients.

We propose a framework using localized kernels based on Chebyshev polynomials to address classification analogously to signal separation. Our approach builds on [6] offering a solid theoretical basis. Though detailed proofs are omitted, extensive experiments highlight the effectiveness of our algorithm, further enhanced by an iterative active learning process across diverse datasets.

II. Related Work

Active learning has been explored extensively to improve learning efficiency. Settles [7] provides a comprehensive survey of active learning strategies, while Dasgupta [8] offers theoretical insights into its effectiveness.

In super-resolution, Candès and Fernandez-Granda [9] introduced convex optimization techniques for recovering point sources, and Tang et al. [10] applied compressed sensing to spectral estimation.

Cloninger and Mhaskar [6] introduced cautious active clustering using localized kernels based on Hermite polynomials. Our approach requires a smaller number of data points.

III. Theoretical Framework

Let μ^* denote a probability measure supported on $\mathbb{X} \subset \mathbb{R}^q$, representing the distribution of a data set. With an appropriate stereographic projection, we may in fact assume that $\mathbb{X} \subset \mathbb{S}^q$, where $\mathbb{S}^q$ denotes the unit sphere of $\mathbb{R}^{q+1}$. We denote by $\mathbb{B}(x,r)$ the spherical cap of radius r, centered at $x \in \mathbb{S}^q$, and for any subset $A \subseteq \mathbb{S}^q$, write $\mathbb{B}(A,r) = \cup_{x\in A}\mathbb{B}(x,r)$. We say that μ^* is **detectable** if there exists $\alpha > 0$ such that for every $x \in \mathbb{S}^q$,

$$\begin{aligned}
&\mu^*(\mathbb{B}(x,r) \leq c_1 r^\alpha, \text{ for } r > 0,\\
&\mu^*(\mathbb{B}(x,r) \geq c_2 r^\alpha, \text{ for } 0 < r \leq 1.
\end{aligned} \tag{1}$$

It is hoped that the support $\mathbb{X}$ should have lower dimension than the ambient dimension q. The first inequality is satisfied, for example, if $\mathbb{X}$ is a submanifold of dimension α and μ^* is its Riemannian volume measure. The second inequality above is analogous to the minimal magnitude of the coefficients in the signal separation problem, and plays the same role in our theory.

We assume that there are finitely many (K) classes in the data set, with the k-th class arising from a probability distribution μ_k. Thus, we assume that μ^* is a convex combination of μ_k's. Ideally, the support of the measures μ_k should be disjoint. To allow for an overlap of class boundaries, we assume instead that for any $\eta > 0$, the support $\mathbb{X}$ of μ^* is a disjoint union of K_η sets $S_{k,\eta}$, separated by a threshold $\eta > 0$ and an extra set, $S_{K_\eta+1}$ representing the overlaps. We assume that $\mu^*(S_{K_\eta+1}) \to 0$ as $\eta \to 0+$. We will say that μ^* has a **fine structure** if it is detectable and such a partition exists. Our goal is to separate these sets.

Towards this goal, we will use the localized polynomial defined on $[-1,1]$ by

$$\Phi_n(\cos\theta) = 1+2\sum_{\ell=1}^{n-1} H\left(\frac{\ell}{n}\right)\cos(\ell\theta), \qquad \theta \in [0,\pi]. \tag{2}$$

where the degree n is a tunable parameter, $H : [0,\infty) \to [0,1]$ is infinitely differentiable and non-increasing function, such that $H(t) = 1$ if $t \in [0,1/2]$ and $H(t) = 0$ if $t \geq 1$. It is easy to prove using the Poisson summation formula [3] that for any $S \geq 2$, there exists a constant $c = c(H,S) > 0$ such that

$$|\Phi_n(\cos\theta)| \leq \frac{cn}{\max(1,(n\theta)^S)}, \qquad \theta \in [0,\pi],\ n \geq 2. \tag{3}$$

We note that if $x, y \in \mathbb{S}^q$, then $\Phi_n(\langle x, y\rangle)$ is a spherical polynomial in both x and y. Since the geodesic distance between x and y is given by $\rho(x, y) = \arccos(\langle x, y \rangle)$, Eqn. (3) becomes

$$|\Phi_n(\langle x, y \rangle)| \leq \frac{cn}{\max(1, (n\rho(x, y))^S)}, \; x, y \in \mathbb{S}^q, \; n \geq 2. \tag{4}$$

IV. Main results

In this section, we introduce the main theorems of this paper. We start by defining our measure estimator:

$$F_{n,M}(x) := \frac{1}{M} \sum_{j=1}^{M} \Phi_n(\langle x, x_j \rangle)^2. \tag{5}$$

We use this estimator to generate sets which approximate the support of μ^*. In particular, we define:

$$\mathcal{G}_n(\Theta) := \left\{ x \in \mathbb{S}^q F_{n,M}(x) \geq \Theta \max_{1 \leq k \leq M} F_{n,M}(x_k) \right\}. \tag{6}$$

The following theorem is proved in [11]

Theorem 1. *Let μ^* be a probability measure with a fine structure given by parameter η and $S \geq q+2$ be an integer. Let $M \geq c_3 n^\alpha \log(n)$ and $\{x_1, x_2, \ldots, x_M\}$ be independent samples from μ^*. There exists $r(\Theta) \sim \Theta^{-1/(S-\alpha)}$ such that with probability at least $1 - c_4/M^{c_5}$ we have*

$$\mathbb{X} \subseteq \mathcal{G}_n(\Theta) \subseteq \mathbb{B}(\mathbb{X}, r(\Theta)/n). \tag{7}$$

Moreover, if $n > 2r(\Theta)/\eta$ there exists a partition $\{\mathcal{G}_{k,\eta,n}(\Theta)\}_{k=1}^{K_\eta+1}$ of $\mathcal{G}_n(\Theta)$ with the following properties:

$$\mathit{dist}(\mathcal{G}_{j,\eta,n}(\Theta), \mathcal{G}_{k,\eta,n}(\Theta)) \geq \eta, \quad j \neq k, \tag{8}$$

and

$$\mathcal{S}_{k,\eta} \subseteq \mathcal{G}_{k,\eta,n}(\Theta) \subseteq \mathbb{B}(\mathcal{S}_{k,\eta}, r(\Theta)/n). \tag{9}$$

Brief Overview of Our Algorithm

The **Salinas** and **Indian Pines** hyperspectral datasets are well-known benchmarks in image analysis, commonly used to assess classification algorithms due to their rich spectral data. The Salinas dataset, captured by the AVIRIS over California's Salinas Valley, provides high spatial resolution with 3.7-meter pixels and 224 spectral bands covering the 0.4–2.5 µm range. It mainly features agricultural areas, making it ideal for land cover and vegetation classification. For our experiments, we use the first 10 classes out of 13, selecting 50% of the data from each class randomly.

The Indian Pines dataset, also from AVIRIS over northwestern Indiana, consists of a 145×145 pixel grid with 220 spectral bands. It poses a classification challenge due to spectral overlap among different land covers. We analyze a 57×41 pixel subset containing corn-notill, stone-steel-towers, woods, soybean-mintil, and grass-trees, with each pixel having 220 spectral features.

Our algorithm, implements Theorem 1 with the following modification. We use an **iterative refinement process** that optimizes clustering by dynamically adjusting kernel parameters and thresholds, validated on real-world datasets like **Salinas** and **Indian Pines**.

We first apply PCA to reduce data from $\mathbb{R}^d$ to $\mathbb{R}^{d'}$ ($d' < d$), preserving maximum variance. The transformed data is normalized and projected onto the unit hypersphere S^q.

Next, we compute the **angle matrix**:

$$A_{ij} = \arccos(\langle x_i, x_j \rangle),$$

quantifying angular distances between $x_i, x_j \in S^q$.

We build an **adjacency graph** $G = (V, E)$ with edges:

$$E = \{(x_i, x_j) : A_{ij} < \eta\},$$

and identify connected components $\{C_{n,\ell}\}$. The refinement process adjusts kernel degree n and threshold Θ to stabilize clustering. Class supports are approximated by $G_n(\Theta)$.

For uncertain components, we iteratively query the most confident points, propagate labels within components, and update n and Θ until labeling stabilizes.

In **post-processing**, we use the **Witness Function Method** as in [12] to propagates labels except instead of Hermite based polynomial kernels in [12] we use the following kernel introduce in [13]: The matrix $\Phi_{n,q}$ is defined as:

$$\Phi_{n,q}(x) = \sum_{k=0}^{n-1} H\left(\frac{k}{n}\right) \frac{P_k^{\left(\frac{q}{2}-1, \frac{q}{2}-1\right)}(1)\, P_k^{\left(\frac{q}{2}-1, \frac{q}{2}-1\right)}(x)}{N_k}$$

where $P_k^{(\alpha,\beta)}(x)$ are Jacobi polynomials with $\alpha = \beta = \dfrac{q}{2} - 1$, and the normalization factor N_k

is:

$$N_k = 2^{\alpha+\beta+1} \frac{\Gamma(k+\alpha+1)\Gamma(k+\beta+1)}{\Gamma(k+1)\Gamma(k+\alpha+\beta+1)} \cdot \frac{1}{2k+\alpha+\beta+1}$$

$$\hat{y}(x) = \arg\max_k \sum_{x_i \in \mathcal{A}_k} \Phi_{n,q}(\langle x, x_i \rangle),$$

where $\mathcal{A}_k$ denotes confidently labeled points of class k.

As shown in Figure 1, our algorithm achieves a success rate of 96.04% using only 3% of the data as queried points. This performance is competitive with state-of-the-art active learning algorithms for the Salinas dataset (see [6]).

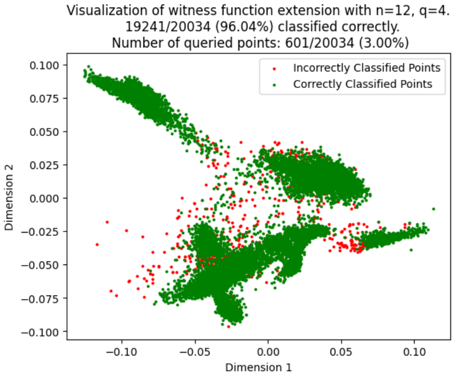

Fig. 1. Salinas dataset.

For the more challenging Indian Pines subset (see Figure 2), our algorithm achieves a success rate of 81.46% using only 7.5% of the data as queried points. This performance is also competitive with state-of-the-art active learning algorithms for the Salinas dataset (see [6]).

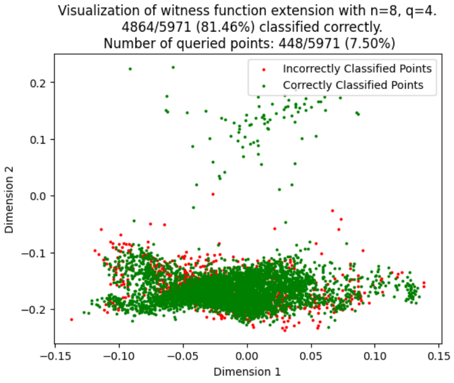

Fig. 2. Indian Pines dataset.

Algorithm 1: Signal Classification via Active learning (SCALe)

```
Input: Dataset $X \subset \mathbb{R}^d$, kernel degree $n$,
       threshold parameter $\Theta$, adjacency
       parameter $\eta$, step size $\eta_{\text{step}}$
Output: Predicted labels $\hat{y}$ for all points
        in $X$
$\mathcal{A} \leftarrow \emptyset$ ;
Apply PCA transformation $\mathbb{R}^d \to \mathbb{R}^{d_{min}}$ ;
Project data onto unit hypersphere $\mathbb{S}^q$ ;
Compute matrix $A_{ij} = \arccos(\langle x_i, x_j \rangle)$ ;
Construct kernel matrix: $\Phi_n(\langle x_i, x_j \rangle)^2$ ;
Prune values from matrix by $\mathcal{G}_n(\Theta)$ ;
while $\eta \leq \eta_{\max}$ do
    Build adjacency graph $G = (V, E)$
      where $E = \{(x_i, x_j) : A_{ij} < \eta\}$ ;
    Identify connected components
      $C_{\eta,\ell}\}_{\ell=1}^{K_n}$ ;
    for $\ell = 1$ to $K_n$ do
        if $C_{\eta,\ell} \cap \mathcal{A} = \emptyset$ then
            $x_i \leftarrow \operatorname{argmax}_{x \in C_{\eta,\ell}} \sum_{j=1}^{M} \Phi_n(\langle x, x_j \rangle)^2$ ;
            $\mathcal{A} \leftarrow \mathcal{A} \cup \{(x_i, f(x_i))\}$ ;
            $\hat{y}(x_j) \leftarrow f(x_i)$ for all $x_j \in C_{\eta,\ell}$ ;
        else if $\forall x_j \in C_{\eta,\ell} \cap \mathcal{A}, f(x_j) = c_\ell$
          then
            $\hat{y}(x_j) \leftarrow c_\ell$ for all $x_j \in C_{\eta,\ell}$ ;
    $\eta \leftarrow \eta + \eta_{\text{step}}$ ;
Identify uncertain points
  $\mathcal{C}_{\text{uncertain}} \leftarrow X \setminus \bigcup_{\eta,\ell} C_{\eta,\ell}$ ;
$\mathcal{A}_k \leftarrow \{x_j : \hat{y}(x_j) = k\}$ ;
foreach $x_j \in \mathcal{C}_{uncertain}$ do
    $\hat{y}(x_j) \leftarrow \operatorname{argmax}_k \sum_{x_i \in \mathcal{A}_k} \Phi_{n,q}(\langle x_j, x_i \rangle)$
return $\hat{y}$
```

V. Conclusion

In this paper, we introduced an active learning algorithm inspired by signal separation principles, demonstrating competitive performance on hyperspectral datasets with minimal labeled data. Our approach effectively identifies class supports even in the presence of overlapping distributions.Future work will focus on evaluating our algorithm's generalizability across datasets from domains like medical imaging, remote sensing, and social networks to assess its adaptability to different classification tasks.

References

[1] E. M. Stein, *Singular integrals and differentiability properties of functions.* Princeton university press, 1970.

[2] H. Mhaskar, S. Kitimoon, and R. G. Raj, “Robust and tractable multidimensional exponential analysis,” *arXiv preprint arXiv:2404.11004*, 2024.

[3] F. Filbir, H. N. Mhaskar, and J. Prestin, “On the problem of parameter estimation in exponential sums,” *Constructive Approximation*, vol. 35, no. 3, pp. 323–343, 2012.

[4] H. N. Mhaskar and J. Prestin, “On the detection of singularities of a periodic function,” *Advances in Computational Mathematics*, vol. 12, no. 2-3, pp. 95–131, 2000.

[5] ——, “On local smoothness classes of periodic functions,” *Journal of Fourier Analysis and Applications*, vol. 11, no. 3, pp. 353–373, 2005.

[6] A. Cloninger and H. N. Mhaskar, “Cautious active clustering,” *Applied and Computational Harmonic Analysis*, vol. 54, pp. 44–74, 2021.

[7] B. Settles, “Active learning literature survey,” 2009.

[8] S. Dasgupta, “Two faces of active learning,” *Theoretical computer science*, vol. 412, no. 19, pp. 1767–1781, 2011.

[9] E. J. Candès and C. Fernandez-Granda, “Towards a mathematical theory of super-resolution,” *Communications on pure and applied Mathematics*, vol. 67, no. 6, pp. 906–956, 2014.

[10] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, “Compressed sensing off the grid,” *IEEE transactions on information theory*, vol. 59, no. 11, pp. 7465–7490, 2013.

[11] H. N. Mhaskar and R. O’Dowd, “Signal separation approach for classification,” in preparation.

[12] H. N. Mhaskar, A. Cloninger, and X. Cheng, “A witness function based construction of discriminative models using hermite polynomials,” *Frontiers in Applied Mathematics and Statistics*, vol. 6, p. 31, 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fams.2020.00031

[13] H. N. Mhaskar and R. O’Dowd, “Learning on manifolds without manifold learning,” *arXiv preprint arXiv:2402.12687*, 2024.