Condensation for the Approximate Nearest-Neighbor Rule

Alejandro Flores-Velazco & David Mount

afloresv@cs.umd.edu

mount@cs.umd.edu

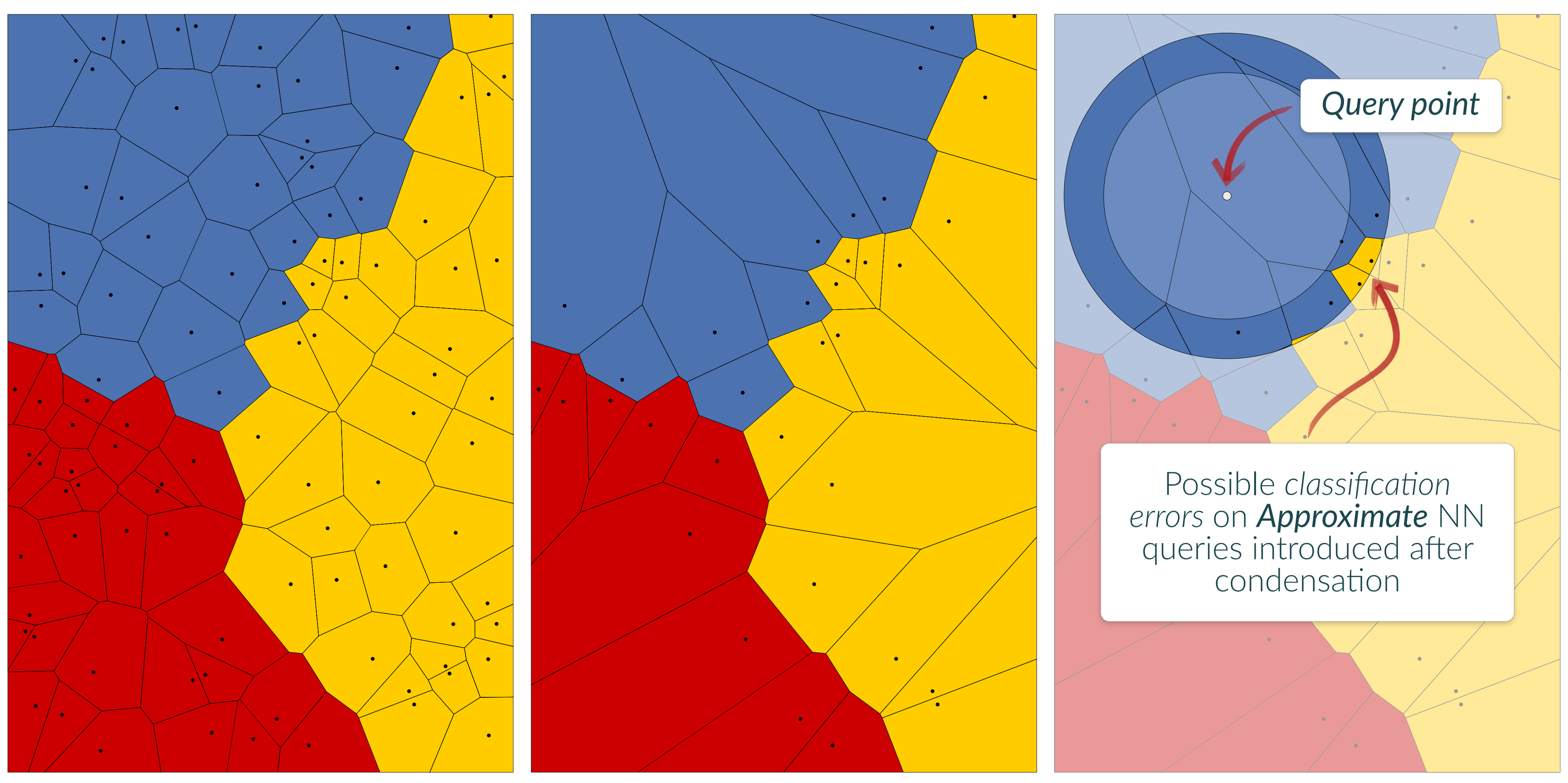
University of Maryland

Department of Computer Science

Consider the problem of

Nearest-Neighbor Condensation

Consider a training set P of labeled points from a metric space $(\mathcal{X}, \mathsf{d})$, to be used by the NN rule to classify new query points. The **NN condensation** problem deals with replacing the training set P with a significantly smaller subset without affecting the classification accuracy under the NN rule.



Training Set — NN Condensation — *Drawbacks!*

The goal of NN *condensation* is to find a subset of P s.t. under the NN rule, every point in P is correctly classified. Such condensed set is called a **consistent** subset of P.

The notion of consistency is defined on ***exact*** NN queries. What if we consider an ***approximate*** version of this? We propose the α-RSS algorithm to select such subsets.

Preliminaries

An ***enemy*** of a point $p \in \mathsf{P}$ is any point in P of a different class. According to the metric ***d***, the ***nearest enemy*** of p is denoted as $\mathsf{NE}(p)$ and its ***NE distance*** as $\mathsf{d}_{\mathrm{ne}}(p)$.

A parameterized algorithm for NN condensation

α-Relaxed Selective Subset

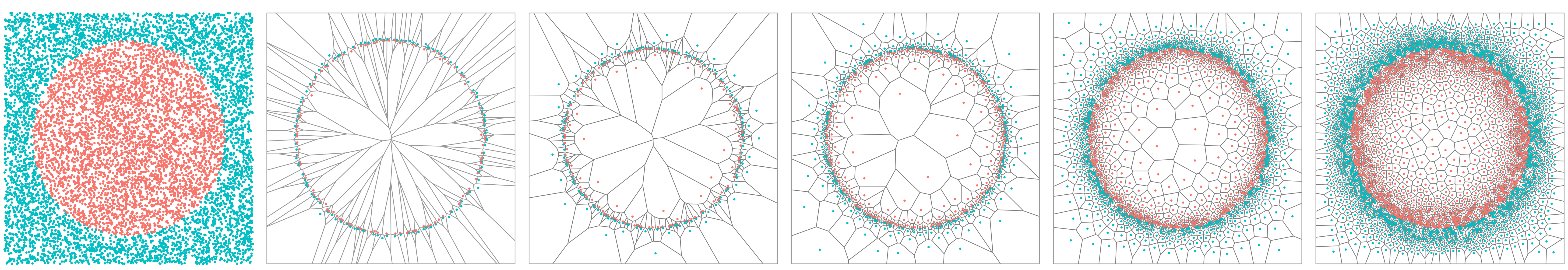
Input: initial training set P, and value $\alpha \geq 0$

Output: condensed training set α-RSS $\subseteq$ P

```
Let {p_i}_{i=1}^n be the points of P sorted in increasing order of NE distance d_ne(p_i)
α-RSS ← ∅
Foreach p_i ∈ P where i = 1...n do
    If ∀ r ∈ α-RSS, (1+α)·d(p_i, r) ≥ d_ne(p_i) then
        α-RSS ← α-RSS ∪ {p_i}
Return α-RSS
```

The α-RSS algorithm computes a consistent subset of P in $\mathcal{O}(n^2)$ worst-case time, and it is order independent.

- Order independence means the resulting subset is not determined by the order in which points are considered by the algorithm.
- Every point in P is correctly classified by α-ANN queries on α-RSS
- 0-RSS equals RSS and ∞-RSS equals P.



Training Set P — 10^6 *points*; 0-RSS — *233 points*; 0.1-RSS — *300 points*; 0.5-RSS — *540 points*; 2-RSS — *1395 points*; 5-RSS — *2894 points*

Results · Guarantees on the...

Classification Accuracy of α-RSS

The ***chromatic density*** $\delta(q)$ of a query point $q \in \mathcal{X}$ is defined as

$$\delta(q, \mathsf{P}) = \frac{\mathsf{d}_{\mathrm{ne}}(q, \mathsf{P})}{\mathsf{d}_{\mathrm{nn}}(q, \mathsf{P})} - 1$$

Where $\mathsf{d}_{\mathrm{nn}}(q, \mathsf{P})$ is the NN distance of q.

Theorem

Consider two parameters $\varepsilon_1 \geq \varepsilon_2 > 0$ both upper bounded by some constant, and set $\alpha = \Omega(1/(\varepsilon_1 - \varepsilon_2))$. Now, if a query point $q \in \mathcal{X}$ has $\delta(q, \mathsf{P}) > \varepsilon_1$ then $\delta(q, \alpha\text{-RSS}) > \varepsilon_2$.

Theorem

The subset of P selected by $(2/\varepsilon)$-RSS is a weak ε-coreset for the chromatic nearest-neighbor of P on query points $q \in \mathcal{X}$ where $\delta(q, \mathsf{P}) > \varepsilon$.

Sufficient conditions for correct classification after NN condensation using α-RSS.

Results · Upper-bounds for...

The size of α-RSS

Let Δ be the ***spread*** of P (i.e., the ratio between the *largest* and *smallest* pairwise distances in P) and κ the ***number of NE*** points in P.

Theorem

Consider $(\mathcal{X}, \mathsf{d})$ to be a **doubling space** with *doubling dimension* $\mathrm{ddim}(\mathcal{X})$, then:

$$|\alpha\text{-RSS}| = \mathcal{O}\left(\kappa\, \alpha^{\mathrm{ddim}(\mathcal{X})+1} \log \Delta\right)$$

Theorem

Consider $(\mathcal{X}, \mathsf{d})$ to be the **Euclidean space**, s.t. $\mathsf{P} \subset \mathbb{R}^d$, then:

$$|\alpha\text{-RSS}| = \mathcal{O}\left(\kappa\, \alpha^{d-1} \log \Delta\right)$$