Nearest Neighbor Condensation with Guarantees

Alejandro Flores-Velazco
Department of Computer Science, University of Maryland, College Park, MD 20742, USA
afloresv@cs.umd.edu

Abstract

Given a training set of labeled points $\mathsf{P} \subset \mathbb{R}^d$, the *Nearest Neighbor* (NN) *rule* classifies a query point with the label of its nearest neighbor in P. The problem of NN *condensation* deals with selecting a subset of P, with the goal of reducing storage and query complexity of the NN rule, while maintaining its original classification accuracy. Even though a significant number of NN condensation algorithms have been proposed, surprisingly, no bounds are known for the amount of reduction that most of these techniques achieve. Moreover, these techniques focus on preserving the classification accuracy on *exact* NN queries, ignoring the effect this condensation might have on *approximate* NN queries.

In this paper, we present theoretical guarantees for state-of-the-art NN condensation techniques. We first consider the MSS algorithm, and show it selects $\mathcal{O}(k)$ points, where k is the number of *border points* of P. These border points are those that define the boundaries between sets of points of different classes, and provide a useful measure of their complexity. Additionally, we propose RSS, a relaxed version of MSS that selects both border and internal points of P. We prove RSS selects $\mathcal{O}(k \log \Delta)$ points, where Δ is the spread of P. Furthermore, assuming query points are distributed uniformly, we show the probability of correctly classifying such query points using ANN queries on RSS, grows exponentially *w.r.t.* the size of RSS.

2012 ACM Subject Classification I.1.2 Algorithms, I.3.5 Computational Geometry and Object Modeling, I.2.6 Learning

Keywords and phrases approximation, nearest neighbor, classification, condensation

1 Introduction

In the context of *non-parametric classification*, a training set P of n labeled points in $\mathbb{R}^d$ is given. Each point $p \in \mathsf{P}$ belongs to a one of set of discrete classes, indicated by its label $l(p)$. Given an unlabeled query point $q \in \mathbb{R}^d$, the goal of a *classifier* is to predict the label of q (i.e., to classify q) using the training set P. The *Nearest Neighbor* (NN) *rule* is one such classification technique: it classifies the query point q with the label of its nearest neighbor in P, that is, $l(\mathsf{NN}(q))$.

Despite its simplicity, the NN rule exhibits good classification accuracy. Theoretical results [10, 4, 5] show that its probability of error is bounded by twice the Bayes probability of error (the minimum of any decision rule). Nonetheless, the NN rule is often criticized on the basis of its memory requirements, as P must be stored to answer queries. Furthermore, the complexity of answering these queries clearly depends on the size and dimensionality of P. Clearly, these drawbacks open an important research question: can P be reduced without affecting the classification accuracy of the NN rule? This problem is called NN Condensation[1].

[1] The problem of NN Condensation is sometimes called Prototype Selection or Instance Selection

1.1 Related work

A natural approach for condensing P would be to consider its Delaunay triangulation. In this context, any point with at least one Delaunay neighbor of different class, is called a *border point*, otherwise is called an *internal point*. Clearly, the border points of P are the ones that completely characterize the boundaries between sets of points of different classes, which are often referred as the *decision boundaries* of P. Therefore, one approach for NN condensation is to select the set of all border points of P, such that its decision boundaries are preserved; this is called *Voronoi condensation* [13]. Unfortunately, a straightforward algorithm would be impractical in high-dimensional spaces. For the planar case, an output-sensitive algorithm was proposed [3] running in $\mathcal{O}(n \log k)$ time, where k is the number of border points of P. Yet, it remains an open problem whether a similar result is possible in higher dimensions.

By relaxing the restriction of preserving the decision boundaries of P, other properties can be exploited to condense P. We describe two such properties, called *consistency* and *selectivity*, which have been widely used in the literature for NN condensation. First, let's introduce a useful concept: an *enemy* of a point $p \in \mathsf{P}$ is said to be any point in P of different class as p. Then, denote the *nearest enemy* (or simply NE) of p as $\mathsf{NE}(p)$, and the NE *distance* as $\mathrm{d}_{\mathsf{NE}}(p) = \mathrm{d}(p, \mathsf{NE}(p))$. Finally, denote the NE *ball* of p as the ball $B(p, \mathrm{d}_{\mathsf{NE}}(p))$.

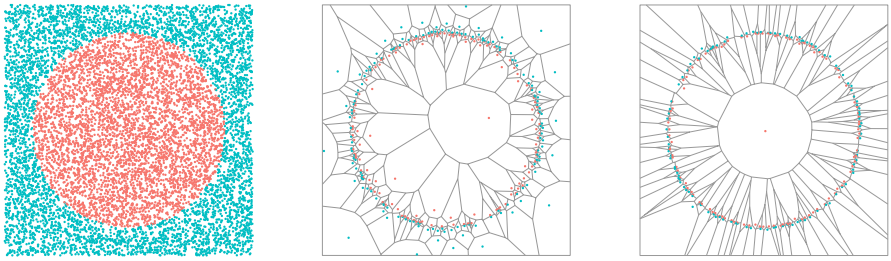
▶ **Definition 1** (Consistency and Selectivity)**.** Let $\mathsf{R} \subseteq \mathsf{P}$ we say that:

- R is a *consistent* subset of P *iff* $\forall\ p \in \mathsf{P}$ its NN in R is closer to p than its NE in R.
- R is a *selective* subset of P *iff* $\forall\ p \in \mathsf{P}$ its NN in R is closer to p than its NE in P.

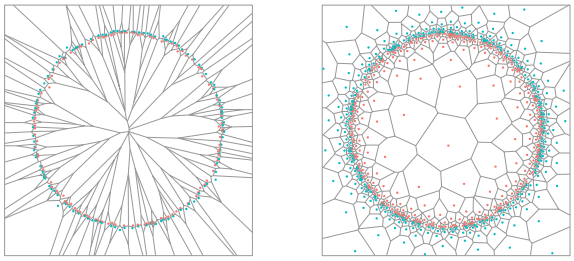
Clearly, selectivity implies consistency, as the NE distance in R of any point is at least its NE distance in P. Moreover, note that the set of all border points, which preserves the decision boundaries of P, is both selective and consistent. Intuitively, a consistent subset R implies that every point 'removed' from P (*i.e.*, every point in the set $\mathsf{P} \setminus \mathsf{R}$) can be correctly classified by NN queries over R. Therefore, while *Voronoi condensation* guarantees the same classification of any query point, before and after condensation, a consistent subset can only guarantee the correct classification of the points removed. In fact, NN condensation is defined as the problem of finding consistent subsets; *i.e.*, it's defined using the weaker property out of the three described.

Unfortunately, it has been shown that the problems of finding minimum-size consistent and selective subsets are both NP-complete [14, 15]. While these results are more recent, almost all research on the problem concentrated in proposing heuristics for finding subsets with these properties (for a comprehensive survey see [11, 12, 8]). Among them, CNN (Condensed Nearest Neighbor) [7] was the first algorithm proposed for computing consistent subsets. Even though it has been widely used in the literature, CNN suffers from several drawbacks: it's running time is cubic in the worst-case, and the resulting subset is order-dependent[2]. Recent efforts resulted in FCNN (Fast CNN) [1] and MSS (Modified Selective Subset) [2], which produce consistent and selective subsets respectively. Both algorithms run in $\mathcal{O}(n^2)$ worst-case time, and are order-independent. For these features, both algorithms are considered the state-of-the-art for the NN condensation problem. Unfortunately, to the best of our knowledge, no bounds are known for the size of the subsets generated by any of these heuristics. More recently, an approximation algorithm called NET [6] was proposed, along with "almost matching" hardness lower bounds. While this proves nearly optimal behavior in

[2] Order-dependence means the resulting subset is determined by the order in which points are considered by the algorithm.



(a) Dataset P (10^4 points) **(b)** CNN (281 points) **(c)** FCNN (222 points)



(d) MSS (233 points) **(e)** RSS (627 points)

Figure 1 An illustrative example of the subsets selected by CNN, FCNN, MSS, and RSS, from an initial dataset P of 10^4 points. While most algorithms focus on selecting border points, or points near the decision boundaries of P, RSS also selects internal points with a selection density relative to the distance to the decision boundaries of P.

the worst-case, the condensation using NET is too strict to be of any practical use. Basically, NET produces an γ-net of P, with γ equal to the minimum NE distance in P. While this subset is consistent, it allows very little room for condensing P (*i.e.*, in general, not many points of P can be covered with γ-balls, and therefore removed).

Algorithm 1: Modified Selective Subset

Input: Initial point set P
Output: Condensed point set MSS $\subseteq$ P

Let $\{p_i\}_{i=1}^{n}$ be the points of P sorted in increasing order of NE distance $\mathrm{d}_{\mathsf{NE}}(p_i)$
MSS $\leftarrow \varnothing$
foreach $p_i \in$ P, *where* $i = 1 \dots n$ **do**
    if $\neg\exists\ r \in$ MSS s.t. $\mathrm{d}(p_i, r) < \mathrm{d}_{\mathsf{NE}}(p_i)$ **then**
        MSS $\leftarrow$ MSS $\cup \{p_i\}$
return MSS

1.2 Drawbacks of NN Condensation

In general, NN condensation algorithms focus on selecting border points or points close to the borders (see Figure 1). These points are the ones that characterize the decision boundaries of P, and therefore, are key in maintaining the classification accuracy of the NN rule after condensation. However, this is only true for *exact* NN queries; we argue that removing internal points reduces actually the accuracy of the NN rule when performing *approximate*

NN queries. If only border points are kept after condensation, a query point that is far from the decision boundaries of P is likely to be misclassified. That is, the $(1+\epsilon)$-ball centered at the query point could contain *enemy* points, and a $(1+\epsilon)$-ANN query can return such enemy point as a valid answer. This notion is formalized in [9] as the *chromatic density* of a query point, and it's defined as $\delta(q) = \frac{\mathrm{d}_{\mathsf{NE}}(q)-\mathrm{d}_{\mathsf{NN}}(q)}{\mathrm{d}_{\mathsf{NN}}(q)}$. It's easy to see how, if $\delta(q) > \epsilon$, q will always be correctly classified by $(1+\epsilon)$-ANN queries. Therefore, by removing internal points, these heuristics can significantly reduce the *chromatic density* of any given query point, and thus, decrease the classification accuracy after condensation, when using ANN queries.

1.3 Contributions

In this paper, we present theoretical guarantees on new and existing heuristic algorithms for the problem of NN condensation. The following is a summary of our results.

- We propose RSS (Relaxed Selective Subset), a new heuristic algorithm for NN condensation, designed to select both border and internal points of P. This algorithm is comparable with other state-of-the-art algorithms (*e.g.*, MSS and FCNN) for the problem.
- Additionally, we provide an upper-bound on the size of RSS. We show RSS selects at most $\mathcal{O}(k \log \Delta)$ points, where k is the number of border points of P, and Δ its spread.
- Similarly, we provide an upper-bound for the size of MSS, showing it selects at most $\mathcal{O}(k)$ points, where again, k is the number of border points of P.
- Assuming query points are drawn uniformly at random from the minimum enclosing ball of P, we show the probability of equally classifying such query points using both *exact* and *approximate* NN queries on RSS, grows exponentially *w.r.t.* the size of RSS.

2 Relaxed Selective Subset

Let's first consider one of the state-of-the-art heuristic algorithms for the problem, known as MSS or Modified Selective Subset (see Algorithm 1). This algorithm is quite simple; the points of P are examined in increasing order *w.r.t.* their NE distance, and any point that doesn't meet the *selective* condition is added to the resulting subset. That is, a point $p \in \mathsf{P}$ is added to MSS *iff* no point in (the current) MSS is closer to p than its NE in P.

Algorithm 2: Relaxed Selective Subset

Input: Initial point set P
Output: Condensed point set $\mathsf{RSS} \subseteq \mathsf{P}$

```
1 Let {p_i}_{i=1}^n be the points of P sorted in increasing order of NE distance d_NE(p_i)
2 RSS ← ∅
3 foreach p_i ∈ P, where i = 1...n do
4     if ¬∃ r ∈ RSS s.t. d(p_i, r) < d_NE(r) then
5         RSS ← RSS ∪ {p_i}
6 return RSS
```

We propose RSS, or Relaxed Modified Subset (see Algorithm 2), as a simple modification on the MSS algorithm. This modification consists on a relaxation of the selective condition; instead of using the NE distance of p_i, we use the one of r. Thus, the idea behind RSS is that every point p_i selected by the algorithm "prunes away" any other point with higher NE distance which is contained in the NE ball of p_i (see Figure 2). Intuitively, this implies that

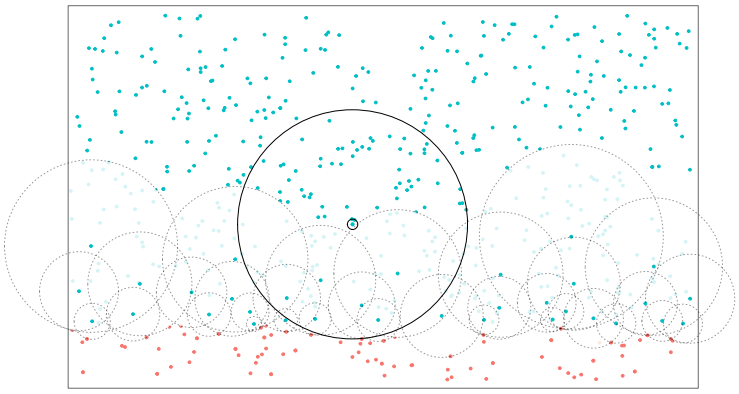


Figure 2 Illustration of the selection process of RSS. The point highlighted is the one being examined by the algorithm. As it isn't contained in the NE balls of any of the previously selected points (represented here as dashed circles), the point is added to RSS. Similarly, this means that any other point **1.** with higher NE distance, and **2.** that is contained in the NE ball of this point, won't be added to RSS.

points close to the decision boundaries of P, and thus, with high NE distance, can prune fewer points than those far from these boundaries. Therefore, the selection of RSS gets sparser according to the distance to the decision boundaries of P.

Just as MSS, the RSS algorithm meets some basic but important properties, which makes it comparable with other state-of-the-art algorithms for NN condensation. These are described in the following theorem.

▶ **Theorem 2.** RSS *is a selective (therefore consistent) subset of* P*, can be computed in worst-case* $\mathcal{O}(n^2)$ *time, and it's order-independent.*

Proof. Let's first show why RSS is selective. Consider any point $p \in \mathsf{P}$; if $p \in \mathsf{RSS}$, it clearly holds the selective property. Otherwise, if $p \notin \mathsf{RSS}$, by construction of RSS there exists a point $r \in \mathsf{RSS}$ *s.t.* $\mathrm{d}(p,r) < \mathrm{d}_{\mathsf{NE}}(r)$. As point r was selected before checking p, we know $\mathrm{d}_{\mathsf{NE}}(r) \leq \mathrm{d}_{\mathsf{NE}}(p)$. Finally, this means there exist a point $r \in \mathsf{RSS}$ *s.t.* $\mathrm{d}(p,r) < \mathrm{d}_{\mathsf{NE}}(p)$, implying that the selective property holds.

Now, we analyze the worst-case time complexity of the algorithm. First, the sorting step requires $\mathcal{O}(n^2)$ time for computing the NE distances of each point, plus $\mathcal{O}(n \log n)$ time for sorting the points. Additionally, the main loop requires to search the NN in RSS for each point in P. Therefore, this requires an additional $\mathcal{O}(n^2)$ time in the worst case. Finally, the worst-case time complexity of the algorithm is $\mathcal{O}(n^2)$. The order-independence follows from always considering the points of P in the same order: increasing *w.r.t.* their NE distance. ◀

3 Upper-bounds on subset size

Needless to say, knowing the amount of reduction achieved by any NN condensation algorithm is crucial. So far, only experimental results have provided insights into this metric, but no theoretical guarantees were known for the state-of-the-art heuristic algorithms. In this section, we analyze this metric, and provide useful upper-bounds for the size of the subsets selected by both MSS and RSS.

As mean of comparison, we use the set of all border points of P, *i.e.*, the subset selected by *Voronoi condensation*. As mentioned before, these border points completely define the

decision boundaries of P, and the number of such points, namely k, is a natural measure for the complexity of these boundaries. Now, we begin our analysis by creating a natural connection between any point in P and a border point.

▶ **Lemma 3.** *The nearest enemy of $p \in \mathsf{P}$ is a border point of P.*

Proof. Consider a ball $\mathcal{B}$ of maximum radius, holding the following properties: **1.** point $\mathsf{NE}(p)$ is on its surface, **2.** its center is in the line segment between p and $\mathsf{NE}(p)$, and **3.** has no point of P in its interior (it's empty). Being maximal, $\mathcal{B}$ has another point $p' \in \mathsf{P}$ on its boundary (see Figure 3). Ball $\mathcal{B}$ can be obtained by a *pivot operation* as described in [3].

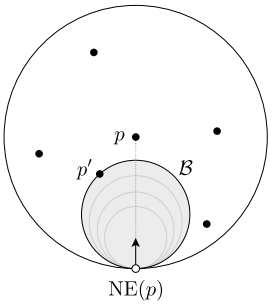


Figure 3 Construction for the proof of Lemma 3.

By construction, $\mathcal{B}$ is completely contained inside the NE ball of p. This implies that p and p' belong to the same class, making p' and $\mathsf{NE}(p)$ enemies. Additionally, by the empty ball property on $\mathcal{B}$, we know that p' and $\mathsf{NE}(p)$ are neighbors in the Delaunay triangulation of P. Therefore, $\mathsf{NE}(p)$ is a border point of P. ◀

This lemma enables the analysis of the sizes of both MSS and RSS. Now, consider the following upper-bound for the size of MSS.

▶ **Theorem 4** (Size of MSS)**.** *Let k be the number of border points of P. Then, $|\mathsf{MSS}| = \mathcal{O}(k)$.*

Proof. Appealing to Lemma 3, the proof follows by a charging argument on each border point. We show that MSS selects a constant number of points for each border point of P.

Consider any border point $p \in \mathsf{P}$, and let MSS_p be the set of points selected by MSS *s.t.* p is their NE. Let $x, y \in \mathsf{MSS}_p$ be two such points. *w.l.o.g.* say that $\mathrm{d}_{\mathsf{NE}}(x) \leq \mathrm{d}_{\mathsf{NE}}(y)$, and by construction of MSS, $\mathrm{d}(x, y) \geq \mathrm{d}_{\mathsf{NE}}(y)$. Thus, consider the triangle $\triangle pxy$. Clearly, the side xy is the larger of such triangle, and thus, the angle $\angle xpy \geq \pi/3$. By a standard packing argument we can upper-bound $|\mathsf{MSS}_p| = \mathcal{O}((3/\pi)^{d-1})$. Therefore, $|\mathsf{MSS}| = \sum_p |\mathsf{MSS}_p| = \mathcal{O}((3/\pi)^{d-1}k)$. Finally, when d is a constant, $|\mathsf{MSS}| = \mathcal{O}(k)$. ◀

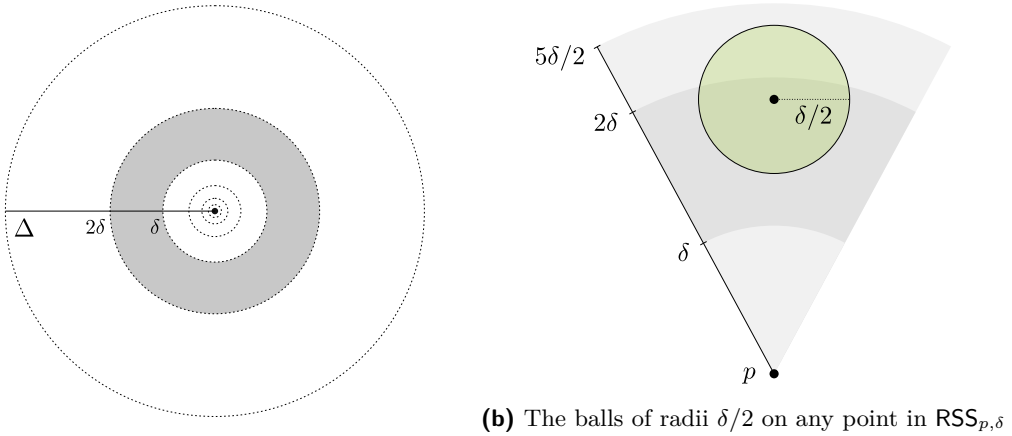
We can now consider a similar analysis for RSS. In order to prove an upper-bound for the size of RSS, we first need to make the following simple observation.

▶ **Observation 5.** *Take any two points $p, p' \in \mathsf{RSS}$, both with NE distance $\geq \alpha$. Then, the balls $B(p, \alpha/2)$ and $B(p', \alpha/2)$ are disjoint.*

Proof. *w.l.o.g.* say that $\mathrm{d}_{\mathsf{NE}}(p) \leq \mathrm{d}_{\mathsf{NE}}(p')$. By construction of RSS, p' was selected (*i.e.*, added to RSS) after p, thus we know $\mathrm{d}(p, p') > \mathrm{d}_{\mathsf{NE}}(p)$ must hold. Therefore, $\mathrm{d}(p, p') > \alpha$, and the balls $B(p, \alpha/2)$ and $B(p', \alpha/2)$ are disjoint. ◀

▶ **Theorem 6** (Size of RSS)**.** *Let k be the number of border points of P, and Δ the spread of P. Then, $|\mathsf{RSS}| \leq \mathcal{O}(k \log \Delta)$.*

Proof. From Lemma 3, we proceed by a charging argument on each border point. The proof follows by showing that RSS selects at most $\mathcal{O}(\log \Delta)$ points for each border point of P.



(a) Partitioning of RSS into subsets $\mathsf{RSS}_{p,\delta}$.

(b) The balls of radii $\delta/2$ on any point in $\mathsf{RSS}_{p,\delta}$ are disjoint, and contained in the ball $B(p, 5\delta/2)$

Figure 4 Proof ideas for bounding the size of RSS.

Consider any border point $p \in \mathsf{P}$, and let $\mathsf{RSS}_{p,\delta}$ be the points selected by RSS *s.t.* their NE is p and their NE distance is in the range $[\delta, 2\delta)$, for a given $\delta \in [1, \Delta]$. From Observation 5, we know the set of balls of radii $\delta/2$ with centers in $\mathsf{RSS}_{p,\delta}$ must be disjoint. Moreover, the ball of radius $5\delta/2$ centered at p contains all such balls (see Figure 4b). Now, it follows from a standard packing argument with the volumes of these balls that $|\mathsf{RSS}_{p,\delta}| \leq 5^d$.

By considering values of $\delta = 2^i$ for $i \in \{0, 1, \ldots, \lceil \log \Delta \rceil\}$, we cover all points of P that can be charged to the border point p (see Figure 4a). Altogether, we get:

$$|\mathsf{RSS}| = \sum_{p} \sum_{i=0}^{\lceil \log \Delta \rceil} |\mathsf{RSS}_{p,2^i}| = \mathcal{O}(5^d k \log \Delta)$$

Finally, when d is a constant, $|\mathsf{RSS}| = \mathcal{O}(k \log \Delta)$. ◀

4 The importance of internal points

While RSS can select more points than MSS, we argue that these extra points are beneficial during classification. The goal of RSS is to select internal points as well, in order to increase the probability of correctly classifying query points using ANN queries. By correct classification, we mean obtaining the same classification when using both exact and approximate NN queries; *i.e.*, using the classification on exact NN queries as the baseline for comparison. This intuition is formalized as follows:

▶ **Theorem 7.** *Let point $q \in \mathbb{R}^d$ be drawn uniformly at random from the minimum enclosing ball of* P*, where* P *has k border points and spread Δ. Then, the probability of equally classifying q with* RSS*, using both exact and $(1+\epsilon)$-approximate* NN *queries, for $\epsilon \leq 2$, is lower-bounded by:*

$$Pr\Big[l(\mathsf{NN}_{\mathsf{RSS}}(q)) = l(\mathsf{ANN}_{\mathsf{RSS}}(q,\epsilon))\Big] \geq \frac{k\big(2^{d\left\lfloor \frac{|\mathsf{RSS}|}{5^d k} \right\rfloor} - 1\big)}{\Delta^d}$$

Proof. This proof follows by bounding the amount of points (*i.e.*, the d-dimensional volume) which are equally classified using exact and ϵ-approximate NN queries on RSS.

Consider a point $p \in \mathsf{RSS}$. We argue the following are sufficient conditions for q to be equally classified using both exact and approximate NN queries on RSS: p is the NN of q in RSS, and $\mathrm{d}(p,q) \leq \frac{\mathrm{d}_{\mathsf{NE}}(p)}{2+\epsilon}$. This inequality implies that $(1+\epsilon)\cdot\mathrm{d}(p,q) \leq \mathrm{d}_{\mathsf{NE}}(p) - \mathrm{d}(p,q)$. By a simple application of the triangle inequality, we know that $\mathrm{d}(p,\mathsf{NE}(q)) - \mathrm{d}(p,q) \leq \mathrm{d}(q,\mathsf{NE}(q)) = \mathrm{d}_{\mathsf{NE}}(q)$. Additionally, we can say that $\mathrm{d}_{\mathsf{NE}}(p) \leq \mathrm{d}(p,\mathsf{NE}(q))$, as we know $\mathsf{NE}(q)$ is also an enemy of p. Combining these inequalities together, we have that $(1+\epsilon)\cdot\mathrm{d}(p,q) \leq \mathrm{d}_{\mathsf{NE}}(q)$, meaning that q is correctly classified by $(1+\epsilon)$-ANN queries on RSS.

Remember that p is the NN of q in RSS *iff* $q \in \mathrm{Vor}(p,\mathsf{RSS})$, *i.e.*, the Voronoi cell of p in RSS. To lower-bound the volume of $\mathrm{Vor}(p,\mathsf{RSS})$, observe that for any other point $p' \in \mathsf{RSS}$, $\mathrm{d}(p,p') \geq \mathrm{d}_{\mathsf{NE}}(p)/2$. Therefore, $B(p,\mathrm{d}_{\mathsf{NE}}(p)/4) \subseteq \mathrm{Vor}(q,\mathsf{RSS})$.

These results imply that, for any point $p \in \mathsf{RSS}$, the following points have p as their NN in RSS, and are correctly classified when using $(1+\epsilon)$-ANN queries on RSS:

$$\mathrm{Vor}(p,\mathsf{RSS}) \;\cap\; B\Big(p,\frac{\mathrm{d}_{\mathsf{NE}}(p)}{2+\epsilon}\Big) \supseteq B\Big(p,\min\Big(\frac{\mathrm{d}_{\mathsf{NE}}(p)}{4},\frac{\mathrm{d}_{\mathsf{NE}}(p)}{2+\epsilon}\Big)\Big) = B\Big(p,\frac{d_{\mathrm{NE}}(p)}{4}\Big) \tag{1}$$

Now, we bound the NE distance of the points in RSS. Let p_i be the i-th point in RSS sorted by their NE distance *w.r.t.* P. Using the bounds obtained for the proof of Theorem 6, we have that $\mathrm{d}_{\mathsf{NE}}(p_i) \geq 2^{\left\lfloor \frac{i-1}{5^d k}\right\rfloor}$. By plugging this into equation 1, we can bound the amount of points, in terms of d-dimensional volume, that are equally classified by exact NN and $(1+\epsilon)$-ANN queries, as follows:

$$\begin{aligned}
\mathit{Vol}(\mathsf{RSS}) &\geq \sum_{i=1}^{|\mathsf{RSS}|} V_d\Big(\mathrm{d}_{\mathsf{NE}}(p_i)\Big) \geq \sum_{i=1}^{|\mathsf{RSS}|} V_d\Big(2^{\left\lfloor \frac{i-1}{5^d k}\right\rfloor}/4\Big) \\
&\geq 5^d k \sum_{j=0}^{\left\lfloor \frac{|\mathsf{RSS}|}{5^d k}\right\rfloor - 1} V_d\Big(2^j/4\Big) = \frac{\pi^{d/2}\ 5^d k}{\Gamma(\frac{d}{2}+1)\ 4^d} \sum_{j=0}^{\left\lfloor \frac{|\mathsf{RSS}|}{5^d k}\right\rfloor - 1} 2^{jd} \\
&= \frac{\pi^{d/2}\ 5^d k\ \left(2^{d\left\lfloor \frac{|\mathsf{RSS}|}{5^d k}\right\rfloor} - 1\right)}{\Gamma(\frac{d}{2}+1)\ 4^d(2^d-1)} \geq \frac{\pi^{d/2}\ k\left(2^{d\left\lfloor \frac{|\mathsf{RSS}|}{5^d k}\right\rfloor} - 1\right)}{\Gamma(\frac{d}{2}+1)\ 2^d}
\end{aligned}$$

Finally, the ratio between $\mathit{Vol}(\mathsf{RSS})$ (*i.e.*, the volume of correctly classified points), and the volume of the entire space out of which the query points are drawn (*i.e.*, the volume of a d-dimensional ball of radius $\Delta/2$), gives a lower-bound for the probability in hand.

$$\begin{aligned}
Pr\Big[l(\mathsf{NN}_{\mathsf{RSS}}(q)) = l(\mathsf{ANN}_{\mathsf{RSS}}(q,\epsilon))\Big] &\geq \frac{\mathit{Vol}(\mathsf{RSS})}{V_d(\Delta/2)} \\
&\geq \frac{\frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)\ 2^d}\ k\left(2^{d\left\lfloor \frac{|\mathsf{RSS}|}{5^d k}\right\rfloor} - 1\right)}{\frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)\ 2^d}\ \Delta^d} \\
&= \frac{k\left(2^{d\left\lfloor \frac{|\mathsf{RSS}|}{5^d k}\right\rfloor} - 1\right)}{\Delta^d}
\end{aligned}$$

◀

Acknowledgments

The author would like to thank Prof. David Mount for his support, and his helpful insights on the analysis of RSS and MSS. Also, thanks to Prof. Emely Arraíz (*Universidad Simón Bolívar*, Venezuela) for suggesting the problem of NN condensation.

References

1 Fabrizio Angiulli. Fast nearest neighbor condensation for large data sets classification. *Knowledge and Data Engineering, IEEE Transactions on*, 19(11):1450–1464, 2007.

2 Ricardo Barandela, Francesc J Ferri, and J Salvador Sánchez. Decision boundary preserving prototype selection for nearest neighbor classification. *International Journal of Pattern Recognition and Artificial Intelligence*, 19(06):787–806, 2005.

3 David Bremner, Erik Demaine, Jeff Erickson, John Iacono, Stefan Langerman, Pat Morin, and Godfried Toussaint. Output-sensitive algorithms for computing nearest-neighbour decision boundaries. In Frank Dehne, Jörg-Rüdiger Sack, and Michiel Smid, editors, *Algorithms and Data Structures: 8th International Workshop, WADS 2003, Ottawa, Ontario, Canada, July 30 - August 1, 2003. Proceedings*, pages 451–461, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg. URL: `https://doi.org/10.1007/978-3-540-45078-8_39`, `doi:10.1007/978-3-540-45078-8_39`.

4 T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theor.*, 13(1):21–27, January 1967. URL: `http://dx.doi.org/10.1109/TIT.1967.1053964`, `doi:10.1109/TIT.1967.1053964`.

5 Luc Devroye. On the inequality of cover and hart in nearest neighbor discrimination. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (1):75–78, 1981.

6 Lee-Ad Gottlieb, Aryeh Kontorovich, and Pinhas Nisnevitch. Near-optimal sample compression for nearest neighbors. In *Advances in Neural Information Processing Systems*, pages 370–378, 2014.

7 P. Hart. The condensed nearest neighbor rule (corresp.). *IEEE Trans. Inf. Theor.*, 14(3):515–516, September 1968. URL: `http://dx.doi.org/10.1109/TIT.1968.1054155`, `doi:10.1109/TIT.1968.1054155`.

8 Norbert Jankowski and Marek Grochowski. Comparison of instances seletion algorithms i. algorithms survey. In *Artificial Intelligence and Soft Computing-ICAISC 2004*, pages 598–603. Springer, 2004.

9 David M. Mount, Nathan S. Netanyahu, Ruth Silverman, and Angela Y. Wu. Chromatic nearest neighbor searching: A query sensitive approach. *Computational Geometry*, 17(3):97 – 119, 2000. URL: `http://www.sciencedirect.com/science/article/pii/S0925772100000213`, `doi:https://doi.org/10.1016/S0925-7721(00)00021-3`.

10 Charles J Stone. Consistent nonparametric regression. *The annals of statistics*, pages 595–620, 1977.

11 Godfried Toussaint. Open problems in geometric methods for instance-based learning. In Jin Akiyama and Mikio Kano, editors, *JCDCG*, volume 2866 of *Lecture Notes in Computer Science*, pages 273–283. Springer, 2002.

12 Godfried Toussaint. Proximity graphs for nearest neighbor decision rules: Recent progress. In *Progress", Proceedings of the 34 th Symposium on the INTERFACE*, pages 17–20, 2002.

13 Godfried T. Toussaint, Binay K. Bhattacharya, and Ronald S. Poulsen. The application of voronoi diagrams to non-parametric decision rules. *Proc. 16th Symposium on Computer Science and Statistics: The Interface*, pages 97–108, 1984.

14 Gordon Wilfong. Nearest neighbor problems. In *Proceedings of the Seventh Annual Symposium on Computational Geometry*, SCG '91, pages 224–233, New York, NY, USA,

1991. ACM. URL: http://doi.acm.org/10.1145/109648.109673, doi:10.1145/109648.109673.

15 A. V. Zukhba. Np-completeness of the problem of prototype selection in the nearest neighbor method. *Pattern Recognit. Image Anal.*, 20(4):484–494, December 2010. URL: http://dx.doi.org/10.1134/S1054661810040097, doi:10.1134/S1054661810040097.