Nearest Neighbor Condensation with Guarantees

Alejandro Flores V.†
David M. Mount‡

Abstract
The problem of nearest-neighbor (NN) condensation deals with reducing the training-set $P$ for the NN-rule classifier. We propose a new NN condensation algorithm called RSS, and prove that selects a subset of size at most $O(k \log \Delta)$, where $k$ is the number of points in the decision borders of $P$, and $\Delta$ its spread. Similarly, we show that a state-of-the-art algorithm called MSS, selects a subset of size at most $O(k)$. To the best of our knowledge, these are the first bounds on the sizes of point sets generated by NN condensation algorithms. Additionally, we proof the probability of correctly classifying a query point with $\epsilon$-ANN using RSS, grows exponentially with respect to the size of the RSS set.

1 Introduction
Consider a training-set $P \subset \mathbb{R}^d$ of $n$ points, where each point $p \in P$ belongs to one of a set of discrete classes, denoted $l(p)$. The goal of nonparametric classification techniques, is to accurately predict the class of new points. Among the most well-known techniques is the nearest-neighbor (NN) rule, which given a query point $q \in \mathbb{R}^d$, assigns it the class of its nearest neighbor in $P$, denoted $NN(q)$.

Despite its simplicity, the NN rule exhibits good classification accuracy. Theoretical results show that its probability of error is bounded by twice the Bayes probability of error (the minimum of any decision rule). However, the NN rule is often criticized on the basis of its memory requirements, as $P$ must be stored to answer queries. For this reason, we consider the problem of Nearest Neighbor Condensation: selecting a subset of $P$ that maintains its classification performance.

2 Related work
Consider the Delaunay triangulation of $P$. Points with at least one neighbor of a different class are called border points, while others are called internal points. One approach for NN condensation is to select the set of border points of $P$. This is called Voronoi condensation [7]. Unfortunately, a straightforward algorithm is impractical in high-dimensional spaces. For the planar case, an output-sensitive algorithm was proposed [3], which runs in $O(n \log k)$ time when $k$ is the number of border points in $P$. Yet, it remains an open problem whether this is possible for higher dimensions.

There are other properties that can be used to condense $P$. First, let’s introduce the concept of enemy: an enemy of a point $p \in P$ is any point in $P$ of different class, and the nearest enemy of $p$ is denoted as NE($p$). Now, let $R \subseteq P$ we say that:

- $R$ is a consistent subset of $P$ iff for every point $p \in P$, its NN in $R$ is of the same class as $p$ (i.e., $p$ is correctly classified by $R$).
- $R$ is a selective subset of $P$ iff for every point $p \in P$, its NN in $R$ is closer to $p$ than is NE in $P$. Clearly, selectiveness implies consistency.

It has been shown that both problems of finding minimum-size consistent and selective subsets are NP-complete [8]. Therefore, many heuristics have been proposed to find subsets with such properties (for a comprehensive survey see [9]). Among them, CNN [4] was the first algorithm proposed for computing consistent subsets. It has been widely used, despite its worst-case cubic running time, and being order-dependent [1]. Recent efforts resulted in FCNN [1] and MSS [2], which produce consistent and selective subsets respectively. Both algorithms run in $O(n^2)$ time, and are order-independent. Unfortunately, to the best of our knowledge, no bounds are known for the size of the subsets generated by any of these algorithms.

Moreover, condensation can introduce problems during classification. In general, these algorithms focus on selecting border points, as they are key in maintaining the classification accuracy on exact NN queries. However, we argue that keeping internal points can be beneficial when performing approximate NN queries, increasing the chances of correct classification, and reducing the query time [5].

3 Relaxed Selective Subset
We propose an algorithm for NN condensation called RSS, or Relaxed Selective Subset (See Algorithm 1).

First let’s introduce some extra notation. Let $d(p,q)$

[1]Order-dependence means the resulting subset is determined by the order in which points are considered by the algorithm.
be the distance between any two points \( p, q \in \mathbb{R}^d \), and \( d_{\text{NE}}(p) \) be the nearest enemy distance of any point \( p \in P \), defined as \( d_{\text{NE}}(p) = d(p, \text{NE}(p)) \).

**Algorithm 1:** Relaxed Selective Subset

| Input: Initial point set \( P \) |
| Output: Condensed point set \( \text{RSS} \subseteq P \) |
| 1 Let \( \{p_i\}_{i=1}^n \) be the points of \( P \) sorted in increasing order of NE distance \( d_{\text{NE}}(p_i) \) |
| 2 RSS \( \leftarrow \) \( \phi \) |
| 3 foreach \( p_i \in P \), where \( i = 1 \ldots n \) do |
| 4 if \( \neg \exists r \in \text{RSS} \) such that \( d(p_i, r) < d_{\text{NE}}(r) \) then |
| 5 RSS \( \leftarrow \) RSS \( \cup \{p_i\} \) |
| 6 return RSS |

RSS runs in \( O(n^2) \) time, and it is order-independent as points are sorted before being considered. Moreover:

**Theorem 1** RSS is a selective subset of \( P \).

This places RSS among state-of-the-art algorithms for NN condensation. However, how do these algorithms compare in terms of the size of their selected subsets? While other algorithms tend to select border points (or points close to the decision borders), the idea behind RSS is to also select internal points following a particular strategy. Figure 1 illustrates the way RSS selects points in comparison with other NN condensation algorithms. In the following theorems, we formalize the bounds on the size of RSS and MSS.

**Theorem 2** Let \( k \) be the number of border points of \( P \), and \( \Delta \) the spread of \( P \). Then, \( |\text{RSS}| \leq O(k \log \Delta) \).

**Theorem 3** Let \( k \) be the number of border points of \( P \). Then, \( |\text{MSS}| \leq O(k) \).

While RSS can select more points than MSS, we argue that these extra points are beneficial during classification. These internal points will increase the probability of correct classification when using \( \varepsilon \)-approximate NN queries. This intuition is formalized as follows:

**Theorem 4** 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 \( \Delta \). Then, the probability of equally classifying \( q \) with RSS, using both exact and \( \varepsilon \)-approximate NN queries (for \( \varepsilon \leq 2 \)), is bounded by:

\[
Pr\left[l(\text{NN}_{\text{RSS}}(q)) = l(\text{ANN}_{\text{RSS}}(\varepsilon, q))\right] \geq \frac{k^2 \Omega\left(\frac{|\text{RSS}|}{\varepsilon\Delta^d}\right)}{\Delta^d}
\]

**References**


