Near-optimal sample compression for nearest neighbors

Lee-Ad Gottlieb[1], Aryeh Kontorovich[2], Pinhas Nisnevitch[1],

[1]Ariel University, Ariel, Israel [2]Ben-Gurion University of the Negev, Beer Sheva, Israel

1. Advantage of proximity-based methods for classification

Cons of kernel → hyperplane separator
- Assumes Euclidean distances
- Natural distances often highly non-Euclidean

Pros of nearest neighbors:
- Simple, classic learning algorithm (early 50's)
- Requires minimal structure
- Immediate extension to multiclass
- Well-understood consistency properties

See:
- "In Defense of Nearest-Neighbor Based Image Classification" (Boiman et al., 2008)
- "Often yields competitive results" (Weinberger and Saul, 2009)

2. Sample condensing/compression for 1-NN?

Consistent condensing
- Input: Sample $S = S_+ \cup S_-$ in a metric space $(\mathcal{X}, \rho)$
- Condensing: A subset $\tilde{S} \subseteq S$
- Consistency: For any point $x \notin \tilde{S}$
  x's nearest neighbor in $\tilde{S}$ has the same label as x

Nearest neighbor condensing problem
- Input: Sample S
- Output: Minimal consistent $\tilde{S} \subset S$
- NP-hard (Wilfong, 1991; Zukhba, 2010)

Heuristic solutions
- Hart (1968) heuristic
  - Init $\tilde{S} := \emptyset$
  - Greedily add misclassified points of S to $\tilde{S}$
  - Runtime: $O(n^3)$
- Other proposed heuristics: Gates (1972); Ritter et al. (1975); Wilson and Martinez (2000)
- Theoretical guarantees: **none**

3. Benefits of sample compression

Sample condensing
- Pro: Reduced memory usage
- Pro: Faster evaluation on new points
- Pro: Improved generalization bounds
- Familiar example: SVM

Occam-type bound (Graepel et al., 2005)
- whp(δ),

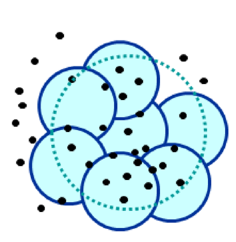
$$\mathrm{err}(h_{\tilde{S}}) \le \frac{1}{n - |\tilde{S}|}\left(|\tilde{S}| \log n + \log n + \log \frac{1}{\delta}\right)$$

holds whenever $\widehat{\mathrm{err}}(h_{\tilde{S}}) = 0$

4. Background to doubling dimension

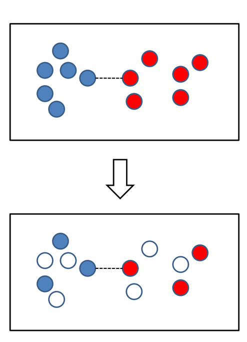
Definition: Doubling dimension
- For any metric space $\mathcal{X}$
- Doubling constant of $\mathcal{X}$:
  Minimum value λ
  such that every big ball B of diameter b
  can be covered by λ small balls with diameter $b/2$
- Doubling dimension of $\mathcal{X}$: $\log \lambda$

History



- Introduced by Assouad (1983)
- Generalizes Euclidean dimension
- Used algorithmically by Clarkson (1999)
  ... and everyone else

Definition: ε-net



- Subset $S' \subset S$
- Packing property: Minimum inter-point distance in S' is ε
- Covering property: Every point in $v \in S$ satisfies $d(v, S') < \varepsilon$
- Construction time: $2^{O(\mathrm{ddim}(S))}|S| \log(1/\varepsilon)$ Krauthgamer and Lee (2004)

5. Main result: Near-optimal sample compression for NN

Define **margin** $\gamma = \mathrm{marg}(S) = \rho(S_+, S_-)$
Our condensing algorithm: Build a γ-net

- **Theorem:** Suppose $\mathrm{diam}(S) = 1$ and $\gamma(S) > 0$.
  Ther exists an algorithm that in time

$$\min\left\{|S|^2, 2^{O(\mathrm{ddim}(S))}|S| \log(1/\gamma)\right\}$$

computes a consistent set $\tilde{S} \subset S$ of size

$$\lceil 1/\gamma \rceil^{\mathrm{ddim}(S)+1}$$

Lower bounds: Algorithm close to best-possible
- **Theorem:** Unless P=NP, cannot approximate S^* within factor $2^{(\mathrm{ddim}(S)\log(1/\gamma))^{1-o(1)}}$
- More precisely:
  1. There exists S with minimal consistent $S^* \subset S$
  2. It is NP-hard to find any consistent set of size

$$|S^*| 2^{(d \log(1/\gamma))^{1-o(1)}}$$

- almost matches upper bound $\lceil 1/\gamma \rceil^{d+1}$

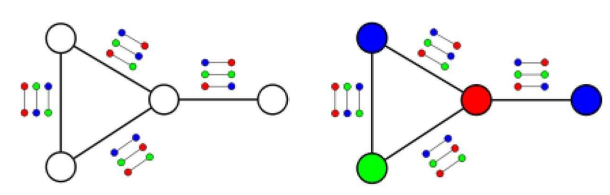
6. Net construction algorithm

Require: S
```
p ← arbitrary point of S
S_1 ← {p}                                          ▷ Top level contains a single point
C(p,0) ← ∅, N(p,0) ← {p}                           ▷ Initialize child and neighbor lists of p
for all q ∈ S do
    P(q,0) ← p                                     ▷ p covers all points
end for
for i = 0, −1, ..., ⌊log γ⌋ + 1 do
    S_{2^{i−1}} ← S_{2^i}                          ▷ All points of level i are present in level i − 1
    for all p ∈ S_{2^{i−1}} do
        C(p, i − 1) ← ∅                            ▷ Initialize child list of p
        for all r ∈ N(p,i) with ρ(p,r) < 4 · 2^{i−1} do
            N(p, i − 1) ← N(p, i − 1) ∪ {r}
            N(r, i − 1) ← N(r, i − 1) ∪ {p}
        end for
    end for
    for all q ∈ S do
        T ← ∪_{r∈N(P(q,i),i)} C(r, i)              ▷ Potential neighbors of q in level i − 1
        if ρ(q,T) < 2^{i−1} then
            P(q, i − 1) ← point r ∈ T with ρ(r,q) < 2^{i−1}
        else
            S_{2^{i−1}} ← S_{2^{i−1}} ∪ {q}        ▷ q is placed in level i − 1
            C(q′, i) ← C(q′, i) ∪ {q}              ▷ Update child list of q's parent
            C(q, i − 1) ← ∅, N(q, i − 1) ← {p}     ▷ Initialize lists of p:
            for all r ∈ T with ρ(q,r) < 4 · 2^{i−1} do
                N(q, i − 1) ← N(q, i − 1) ∪ {r}
                N(r, i − 1) ← N(r, i − 1) ∪ {q}
            end for
        end if
    end for
end for
```

7. Hardness lower bound

NP-hard to find a better compressed set
Reduction:
- From Label Cover problem,
- Reduction holds for Euclidean sets too

Label cover:



- Input: Graph, set of valid labels
- Output: Valid labelling
- Minimization version: Can use multiple colors, minimize number of labels
- Dinur and Safra (2004) showed NP-hard to approximate within a factor $2^{(\log n)^{1-o(1)}}$

8. Empirical results

Experiments
- Data sets: UCI Machine Learning Repository
- Metric: ℓ_1-norm

Table
- (i) Initial sample set size,
- (ii) Percentage of points retained in net extraction

Data set	Original sample	% in net
Skin Segmentation	10000	35.10
Statlog Shuttle	2000	65.75
Covertype 1 vs. 4	2000	35.85
Covertype 4 vs. 6	2000	96.50
Covertype 4 vs. 7	2000	4.40