CAPTCHA: Completely Automated Public Turing test to tell Computers and Humans Apart

Also known as

- Reverse Turing Test
- Human Interactive Proofs

[von Ahn et al., CMU 2000]

Exploit limitations in accuracy of machine pattern recognition

[Diagram showing a CAPTCHA example with 'overlooks' and 'inquiry']

[Diagram showing a scanned type with 'This aged portion of society were distinguished from' and OCR reads as 'niis aged pntkm at society were distinguished frow.']
Amazon’s Mechanical Turk

- “Crowd-sourcing” tedious human intelligence (pattern recognition) tasks
- Which ones are doable by machines?
How Machines Learn to Recognize Patterns

- Sentiment Classification

![Diagram showing sentiment classification with positive and negative categories]

- Sentiment Classification

![Diagram showing sentiment classification with positive and negative categories]
Complexity of Class Boundaries Causes Difficulties in Classification

- Kolmogorov complexity
- A trivial description of a boundary is to list all points & class labels
- Boundary length can be exponential in dimensionality

- Is there a shorter description?
Parameterization of Data Complexity
### Some Useful Measures of Geometric Complexity

<table>
<thead>
<tr>
<th>Degree of Linear Separability</th>
<th>Fisher’s Discriminant Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Find separating hyper-plane by linear programming</td>
<td>Classical measure of class separability</td>
</tr>
<tr>
<td>Error counts and distances to plane measure separability</td>
<td>$f = \frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}$</td>
</tr>
<tr>
<td></td>
<td>Maximize over all features to find the most discriminating</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Length of Class Boundary</th>
<th>Shapes of Class Manifolds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute minimum spanning tree</td>
<td>Cover same-class pts with maximal balls</td>
</tr>
<tr>
<td>Count class-crossing edges</td>
<td>Ball counts describe shape of class manifold</td>
</tr>
</tbody>
</table>
Continuous Distributions in Complexity Space

Real-World Data Sets:
Benchmarking data from UC-Irvine archive
844 two-class problems
452 are linearly separable, 392 non-separable

Synthetic Data Sets:
Random labeling of randomly located points
100 problems in 1-100 dimensions
# Measures of Geometrical Complexity

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>maximum Fisher’s discriminant ratio</td>
</tr>
<tr>
<td>F2</td>
<td>volume of overlap region</td>
</tr>
<tr>
<td>F3</td>
<td>maximum (individual) feature efficiency</td>
</tr>
<tr>
<td>L1</td>
<td>minimized error by linear programming (LP)</td>
</tr>
<tr>
<td>L2</td>
<td>error rate of linear classifier by LP</td>
</tr>
<tr>
<td>L3</td>
<td>nonlinearity of linear classifier by LP</td>
</tr>
<tr>
<td>N1</td>
<td>fraction of points on boundary (MST method)</td>
</tr>
<tr>
<td>N2</td>
<td>ratio of average intra/inter class NN distance</td>
</tr>
<tr>
<td>N3</td>
<td>error rate of 1NN classifier</td>
</tr>
<tr>
<td>N4</td>
<td>nonlinearity of 1NN classifier</td>
</tr>
<tr>
<td>T1</td>
<td>fraction of points with associated adherence subsets retained</td>
</tr>
<tr>
<td>T2</td>
<td>average number of points per dimension</td>
</tr>
</tbody>
</table>
The First 6 Principal Components

<table>
<thead>
<tr>
<th>Component</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prop. of Var.</td>
<td>0.5033</td>
<td>0.1162</td>
<td>0.1064</td>
<td>0.0859</td>
<td>0.0761</td>
<td>0.0521</td>
</tr>
<tr>
<td>Cum. Prop.</td>
<td>0.5033</td>
<td>0.6195</td>
<td>0.7259</td>
<td>0.8118</td>
<td>0.8879</td>
<td>0.9400</td>
</tr>
<tr>
<td>Loadings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td>0.01</td>
<td>0.26</td>
<td>0.03</td>
<td>0.86</td>
<td>-0.26</td>
<td>-0.33</td>
</tr>
<tr>
<td>F2</td>
<td>0.33</td>
<td>0.08</td>
<td>-0.43</td>
<td>-0.12</td>
<td>-0.09</td>
<td>-0.20</td>
</tr>
<tr>
<td>F3</td>
<td>-0.29</td>
<td>0.42</td>
<td>0.03</td>
<td>-0.11</td>
<td>-0.32</td>
<td>0.29</td>
</tr>
<tr>
<td>L1</td>
<td>0.17</td>
<td>0.08</td>
<td>0.68</td>
<td>-0.15</td>
<td>0.00</td>
<td>-0.36</td>
</tr>
<tr>
<td>L2</td>
<td>0.38</td>
<td>0.04</td>
<td>-0.15</td>
<td>-0.14</td>
<td>-0.10</td>
<td>-0.24</td>
</tr>
<tr>
<td>L3</td>
<td>0.38</td>
<td>0.05</td>
<td>-0.16</td>
<td>-0.14</td>
<td>-0.12</td>
<td>-0.23</td>
</tr>
<tr>
<td>N1</td>
<td>0.36</td>
<td>0.30</td>
<td>0.04</td>
<td>0.01</td>
<td>-0.04</td>
<td>0.36</td>
</tr>
<tr>
<td>N2</td>
<td>0.37</td>
<td>-0.02</td>
<td>0.02</td>
<td>0.03</td>
<td>0.12</td>
<td>-0.03</td>
</tr>
<tr>
<td>N3</td>
<td>0.32</td>
<td>0.36</td>
<td>0.00</td>
<td>0.11</td>
<td>-0.03</td>
<td>0.49</td>
</tr>
<tr>
<td>N4</td>
<td>0.24</td>
<td>-0.20</td>
<td>0.52</td>
<td>-0.04</td>
<td>-0.35</td>
<td>0.16</td>
</tr>
<tr>
<td>T1</td>
<td>0.23</td>
<td>-0.32</td>
<td>0.07</td>
<td>0.37</td>
<td>0.57</td>
<td>0.28</td>
</tr>
<tr>
<td>T2</td>
<td>0.08</td>
<td>-0.61</td>
<td>-0.15</td>
<td>0.13</td>
<td>-0.58</td>
<td>0.22</td>
</tr>
</tbody>
</table>
Interpretation of the First 4 PCs

PC 1: 50% of variance: Linearity of boundary and proximity of opposite class neighbor

PC 2: 12% of variance: Balance between within-class scatter and between-class distance

PC 3: 11% of variance: Concentration & orientation of intrusion into opposite class

PC 4: 9% of variance: Within-class scatter
Problem Distribution in 1\textsuperscript{st} & 2\textsuperscript{nd} Principal Components

- Continuous distribution
- Problems known to be easy or difficult occupy opposite ends
- Few outliers
- Empty regions
Apparent vs. True Complexity: Uncertainty in Measures due to Sampling Density

Problem may appear deceptively simple or complex with small samples.
Observations

• Problems distribute in a continuum in complexity space

• Several key measures/dimensions provide independent characterization

• Need further analysis on uncertainty in complexity estimates due to small sample size effects
Relating Classifier Behavior to Data Complexity
Class Boundaries Inferred by Different Classifiers

XCS: a genetic algorithm

Nearest neighbor classifier

Linear classifier
Accuracy Depends on the Goodness of Match between Classifiers and Problems

Problem A

- XCS, error = 1.9%
- NN, error = 0.06%

Problem B

- XCS, error = 0.6%
- NN, error = 0.7%

Better! Better!
Domains of Competence of Classifiers

Given a classification problem, we want to determine which classifier is the best for it.

Can data complexity give us a hint?

Decision Forest

XCS

LC

NN

Complexity metric 1

Metric 2

Here is my problem!
Uncertainty of Estimates at Two Levels

Sparse training data in each problem & complex geometry cause ill-posedness of class boundaries

(uncertainty in feature space)

Sparse sample of problems causes difficulty in identifying regions of dominant competence

(uncertainty in complexity space)
The Landscape Contest in ICPR 2010

Seed Data
(a,b,c,d + pima from UCI archive)

Evaluation Data Used in Contest

Complexity-targeted perturbations

* Golden Datasets
Using cross-validation errors as meta-features, we found a case where meta-learning improves over simple cross-validation for classifier selection.
Complexity and Data Dimensionality: 
Class Separability after Dimensionality Reduction

Feature selection/transformation may change the difficulty of a classification problem:

- Widening the gap between classes
- Compressing the discriminatory information
- Removing irrelevant dimensions

It is often unclear to what extent these happen. We seek quantitative description of such changes.
Spread of classification accuracy and geometrical complexity due to forward feature selection.
Seeking Optimizations Upstream

Back to the application context:
- Use data complexity measures for guidance
- Change the setup, definition of the classification problem
- Collect more samples, in finer resolution, extract more features ...
- Alternative representations:
  - dissimilarity-based? [Pekalska & Duin 2005]

Data complexity gives an operational definition of learnability

Optimization in the upstream: formalize the intuition of seeking invariance, systematically optimize the problem setup and data acquisition scenario to reduce data complexity
Summary

Automatic classification is useful, but can be very difficult. We know the key steps and many promising methods. But we have not fully understood how they work, what else is needed.

We found measures for geometric complexity that are useful to characterize difficulties of classification problems and classifier domains of competence.

Better understanding of how data and classifiers interact can guide practice, and re-establish the linkage between context and solution.