Emerging Work in Open Set Recognition for Vision and Language

Walter J. Scheirer
School of Engineering and Applied Sciences,
Department of Molecular and Cellular Biology, and
Center for Brain Science, Harvard University
Part 1: The Open Set Recognition Problem
Benchmarks in computer vision

Assume we have examples from all classes:

- airplanes
- elephant
- soccer ball
- car
- water lily

Caltech 256
Out in the real world…

Detect the cars in this image while rejecting the trees, signs, telephone poles…
“All positive examples are alike; each negative example is negative in its own way”

Zhao and Huang (with some help from Tolstoy)
CVPR 2001
What is the general recognition problem?

• Duin and Pekalska*: how one should approach multi-class recognition is still an open issue
  - Is it a series of binary classifications?
  - Is it a search performed for each possible class?
  - What happens when some classes are ill-sampled, not sampled at all or undefined?

Vision problems in order of “openness”
Let’s formalize openness

$$\text{openness} = 1 - \sqrt{\frac{2 \times |\text{training classes}|}{|\text{testing classes}| + |\text{target classes}|}}$$
## Examples of openness values

<table>
<thead>
<tr>
<th></th>
<th>Targets</th>
<th>Training</th>
<th>Testing</th>
<th>Openness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typical Multi-class</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>0%</td>
</tr>
<tr>
<td>Face Verification</td>
<td>12</td>
<td>12</td>
<td>50</td>
<td>38%</td>
</tr>
<tr>
<td>Typical Detection</td>
<td>1</td>
<td>100,000</td>
<td>1,000,000</td>
<td>55%</td>
</tr>
<tr>
<td>Object Recognition</td>
<td>88</td>
<td>12</td>
<td>88</td>
<td>63%</td>
</tr>
<tr>
<td>Object Recognition</td>
<td>88</td>
<td>6</td>
<td>88</td>
<td>74%</td>
</tr>
<tr>
<td>Object Recognition</td>
<td>212</td>
<td>6</td>
<td>212</td>
<td>83%</td>
</tr>
</tbody>
</table>
Fundamental multi-class recognition problem

\[
\arg\min_f \left\{ R_\mathcal{I}(f) := \int_{\mathbb{R}^d \times \mathbb{N}} L(x, y, f(x)) P(x, y) \right\}
\]

Ideal Risk  Loss Function  Joint Distribution

Undefined for open set recognition!

Open Space

Negatives

Positives

Specialization
Open Space

- Open space is the space far from known data
- We need to address the infinite half-space problem of linear classifiers
- Principle of Indifference*
  - If there is no known reason to assign probability, alternatives should be given equal probability
  - One problem: we need the distribution to integrate to 1!

Open Space Risk

Open Space Risk: the relative measure of open space to the full space

$$R_\mathcal{O}(f) = \frac{\int_{\mathcal{O}} f(x) dx}{\int_{S_\mathcal{O}} f(x) dx}$$

open space

Open space + positive training examples
The open set recognition problem

Preliminaries

Space of positive class data: \( P \)

Space of other known class data: \( K \)

Positive training data: \( \hat{V} = \{v_1, \ldots, v_m\} \) from \( P \)

Negative training data: \( \hat{K} = \{k_1, \ldots, k_n\} \) from \( K \)

Unknown negatives appearing in testing: \( U \)

Testing data: \( \mathcal{T} = \{t_1, \ldots, t_z\}, t_i \in P \cup K \cup U \)

Assume the problem openness is \( > 0 \)
The open set recognition problem

Minimize open set risk:

$$\arg\min_{f \in \mathcal{H}} \left\{ R_O(f) + \lambda_r R_E(f(\hat{V} \cup \hat{K})) \right\}$$

- Open Space Risk Associated with $\mathcal{U}$
- Regularization Constant
- Empirical Risk Function
What options do we have to solve this problem?
Binary Classification
Multi-class 1-vs-All Classification
1-class Classification

Why didn’t the 1-class SVM catch on?

- Zhou and Huang *Multimedia Systems* 2003
  - Kernel and parameter selection
    - Gaussian kernels lead to over-fitting
    - Parameters chosen in *ad hoc* fashion
    - An issue in other domains too!

Other approaches


Let’s include open space risk in our optimization problem
Slab Model

Negatives

Positives

Ω

A
Base Linear 1-vs-Set Machine
Generalization
Specialization
Open space risk for linear slab model

\[ \delta_A \quad \text{Marginal distance of near plane} \]

\[ \delta_{\Omega} \quad \text{Marginal distance of far plane} \]

\[ \delta^+ \quad \text{Separation needed to account for all positive data} \]

\[ \frac{\delta_{\Omega} - \delta_A}{\delta^+} \quad \text{Overgeneralization risk} \]

\[ \frac{\delta^+}{\delta_{\Omega} - \delta_A} \quad \text{Overspecialization risk} \]
Open space risk for linear slab model

\[ R_\varsigma = \frac{\delta_\Omega - \delta_A}{\delta^+} + \frac{\delta^+}{\delta_\Omega - \delta_A} + p_A \omega_A + p_\Omega \omega_\Omega \]

Two additional terms

Importance of open space around \( A \)
Importance of open space around \( \Omega \)

Margin around \( A \)
Margin around \( \Omega \)
Sketch of the 1-vs-Set Machine training algorithm

1. Train a linear SVM $f$ using $\hat{V}$ and $\hat{K}$

2. Generate decision scores for each training point in $\hat{V}$ and $\hat{K}$

3. Sort decision scores, where $s_k$ is the minimum and $s_j$ is the maximum

4. Initialize $A$ to margin plane of $f$, and $\Omega$ to $s_j$

5. Iteratively move $A$ to $s_{k+1}$ or $s_{k-1}$, $\Omega$ to $s_{j-1}$ or $s_{j+1}$ to minimize $R_{R_s}(f) + \lambda_r R_\epsilon$
1-vs-Set Machine Plane Refinement

Positive Pressure $p_A > 0$

Plane $A$ after initial optimization

Negative Pressure $p_A < 0$

Plane $A$ after refinement with $p_A = -0.5$
function PREDICT\( (t_x, f, A, \Omega) \)
  if \( A \leq f(t_x) \) and \( f(t_x) \leq \Omega \) then Return +1
  else Return -1
end if
end function
How can we evaluate open set recognition in a controlled manner?
Open Set Object Recognition

Cross-data set methodology*

Training: Caltech 256

Testing: Caltech 256 + ImageNet

Open Universe of 88 classes: 1 positive class, \( n \) training classes, 87 negative testing classes (532,400 images)

Open Universe of 212 classes: 1 positive class, \( n \) training classes, 211 negative testing classes (13,610,400 images)

Features

Histogram of Oriented Gradients


LBP-like descriptor

## 1-vs-Set Machine vs. Typical SVMs

<table>
<thead>
<tr>
<th></th>
<th>2-tailed paired t-test</th>
<th>binary 1-vs-Set</th>
<th>binary linear</th>
<th>binary RBF</th>
<th>1-class 1-vs-Set</th>
<th>1-class linear</th>
<th>1-class RBF</th>
</tr>
</thead>
<tbody>
<tr>
<td>binary 1-vs-Set (HOG 88)</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>binary linear (HOG 88)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>binary RBF (HOG 88)</td>
<td>—</td>
<td>++</td>
<td>—</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>1-class 1-vs-Set (HOG 88)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>**</td>
<td>—</td>
<td>**</td>
<td>—</td>
</tr>
<tr>
<td>1-class linear (HOG 88)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>1-class RBF (HOG 88)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>++</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>binary 1-vs-Set (HOG 212)</td>
<td>—</td>
<td>**</td>
<td>*</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>1-class 1-vs-Set (HOG 212)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>*</td>
</tr>
<tr>
<td>binary 1-vs-Set (LBP-like 88)</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>1-class 1-vs-Set (LBP-like 88)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>**</td>
<td>—</td>
<td>**</td>
<td>—</td>
</tr>
<tr>
<td>binary 1-vs-Set (LBP-like 212)</td>
<td>*</td>
<td>—</td>
<td>—</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>1-class 1-vs-Set (LBP-like 212)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

** 1-vs-Set Machine is statistically significant at p < 0.01

* 1-vs-Set Machine is statistically significant at p < 0.05

++ Baseline Machine is statistically significant at p < 0.01

— No statistical significance
Top 25 classes for the open universe of 88 classes
F-measure as a function of openness

![Graph showing F-measure as a function of openness](image)
Near and far plane pressures for open universe of 88 classes

The second plane has an impact on recognition performance.
Open Set Face Verification

Labeled Faces in the Wild

Genuine Pair

Impostor Pair

Impostor Pair

Impostor Pair

Gallery classes: 12 people with at least 50 images
Impostor classes: 82 other people in LFW
1,316 test images across all classes

Features: LBP-like and Gabor*

Open set face verification

![Graph showing F-measure against openness for different methods: Binary 1-vs-Set Machine, LBP-like, Binary SVM, LBP-like, Binary 1-vs-Set Machine, Gabor, and Binary SVM, Gabor. The graph illustrates how the F-measure decreases with increasing openness.]
Conclusions

Classification assumptions need to be questioned for open set recognition.

Open Space Risk helps us quantify the risk associated with the unknown.

Reducing open space leads to more stable solutions when lots of unknown data appears during testing.
Part 2: Open Set Influence Recognition
We find open set problems in NLP too

An emerging area: **Stylometry**

Machine learning is now able to detect quantifiable style markers in written language.

Common applications: authorship attribution, genre tagging, **textual reuse**

No text exists in isolation: we cannot make *a priori* assumptions on characteristics of style.
Intertextuality in Literature

Kristeva: “Any text is constructed as a mosaic of quotations; any text is the absorption and transformation of another.”

The nature of textual reuse is widely varied:

- Direct quotations
- Loose lexical correspondance
- Idea reuse
- Sound reuse
Quantitative Intertextuality

Since the problem is one of pattern recognition, it is a good candidate for automated assistance by computers.

**Quantitative Intertextuality** is the algorithmic study of information reuse in any semiotic system.

Applications:

- Scholarly work (*e.g.* digital humanities)
- Practical applications (*e.g.* digital forensics)
Case Study: Paul the Deacon’s *Angustae Vitae*

Paul the Deacon: 8th century monk, intellectual, and court poet of the Lombards.

Catullus: influential Latin poet of the late Roman Republic.

*Angustae Vitae* is an epistolary poem that juxtaposes poetry in the classical world with poetry in a Christian monastic context.

Clear Neoteric influence, and surprisingly reminiscent of Catullus.

The Clues

Angustae Vitae is peppered with classical intertexts, but it remains an open question as to whether Paul the Deacon had actually read Catullus.

But we know…

The diction and thematic models recall the Neoterics.

Paul the Deacon was well versed in the poetry of Horace, Ovid and Vergil.

Content suggests familiarity with Catullus 1, 2, 50 and 68.
The Clues

Catullus 2
PASSER, deliciae meae puellae,
quicum ludere, quem in sinu tenere,
cui primum digitum dare appetenti
et acri solet incitare morsus
cum desiderio meo nitenti
carum nescio quid lubet iocari,
credo ut, cum gravis acquiescet ardor,
sit solaciolum sui dolaris,
tecum ludere sicut ipsa possem
et tristis animi levare curas!

Angustae Vitae, lines 1-4:
Angustae vitae fugiunt consortia Musae,
Claustrorum septis nec habitare volunt,
Per rosulenta magis cupiunt sed ludere prata,
Pauperiem fugiunt, deliciasque colunt:
Feature: the functional n-gram

A functional bi-gram is an n-gram-based feature that describes frequently appearing information.

Functional n-grams for sound are:

- Character-level features
- Stand-ins for phonemes
- Similar to function words
- Elements of most of the lexicon

\[
P(e_n|e_{n-N+1}^{n-1}) = \frac{C(e_{n-N+1}^{n-1}e_n)}{C(e_{n-N+1}^{n-1})} \iff freq(e_{n-N+1}^{n-1}e_n) > \phi
\]

The functional n-gram process

Select $x$ of the most frequently occurring n-grams in a sample:

<table>
<thead>
<tr>
<th></th>
<th>804</th>
<th>er</th>
<th>560</th>
<th>ti</th>
</tr>
</thead>
<tbody>
<tr>
<td>778</td>
<td>726</td>
<td>723</td>
<td>re</td>
<td>709</td>
</tr>
<tr>
<td>685</td>
<td>651</td>
<td>615</td>
<td>um</td>
<td>604</td>
</tr>
<tr>
<td>574</td>
<td>454</td>
<td>452</td>
<td>am</td>
<td>468</td>
</tr>
</tbody>
</table>

Compute the probability features:

\[
\frac{\text{Count}(\text{“er”})}{\text{Count}(\text{“e”})} = 0.179785
\]

\[
\frac{\text{Count}(\text{“re”})}{\text{Count}(\text{“r”})} = 0.275447
\]
Open Set Influence Analysis

We want to test the stylistic similarity of any poet to Catullus.

Methodology: train a 1-class SVM on representative samples from Catullus.

Recall that 1-class SVMs tend to overfit the training data.

We don’t need to generalize - we are only interested in samples that fall within the support of the training data.
Open Set Influence Analysis
Low-Probability Analysis

Words that occur infrequently are often intertexts. Incorporate this into the influence analysis process:

Fix a desired probability range for words that occur infrequently.

Scan for n-gram sequences composed of only those words in a particular passage (ignoring all others).

Use as an additional feature to augment our vector of functional n-grams.

\[(P_{low} < \Pr(\text{word}_1) < P_{high}) \ldots (P_{low} < \Pr(\text{word}_2) < P_{high}) \ldots (P_{low} < \Pr(\text{word}_n) < P_{high})\]
Low-Probability Analysis

Key n-gram sequences common to *Angustae Vitae* and Catullus:

- delic(ias|iae) ludure
- flagra(t|ns|bat) amor_
- redde miser poema
Evidence of Catullan Influence in *Angustae Vitae*

Identifying a functional n-gram of interest:

Looked at most frequently occurring character-level bigrams in Catullus 1 - 64.

<table>
<thead>
<tr>
<th>Rank</th>
<th>bi-gram</th>
<th>Probability</th>
<th>Rank</th>
<th>bi-gram</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>er</td>
<td>0.180</td>
<td>6</td>
<td>te</td>
<td>0.251</td>
</tr>
<tr>
<td>2</td>
<td>qu</td>
<td>1.000</td>
<td>7</td>
<td>es</td>
<td>0.146</td>
</tr>
<tr>
<td>3</td>
<td>is</td>
<td>0.169</td>
<td>8</td>
<td>um</td>
<td>0.164</td>
</tr>
<tr>
<td>4</td>
<td>en</td>
<td>0.162</td>
<td>9</td>
<td>in</td>
<td>0.140</td>
</tr>
<tr>
<td>5</td>
<td>re</td>
<td><strong>0.275</strong></td>
<td>10</td>
<td>it</td>
<td>0.133</td>
</tr>
</tbody>
</table>
Evidence of Catullan Influence in *Angustae Vitae*

Sample ‘re’ training data for the poems of Catullus (features most similar to poems 1 & 2):

<table>
<thead>
<tr>
<th>Feature</th>
<th>Poems</th>
<th>Feature</th>
<th>Poems</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.458</td>
<td>1 and 2</td>
<td>0.435</td>
<td>38 and 39</td>
</tr>
<tr>
<td>0.412</td>
<td>2b and 3</td>
<td>0.480</td>
<td>44 and 45</td>
</tr>
<tr>
<td>0.455</td>
<td>4 and 5</td>
<td>0.480</td>
<td>50 and 51</td>
</tr>
<tr>
<td>0.524</td>
<td>7 and 8</td>
<td>0.444</td>
<td>62</td>
</tr>
<tr>
<td>0.500</td>
<td>13</td>
<td>0.463</td>
<td>64</td>
</tr>
<tr>
<td>0.406</td>
<td>17 and 21</td>
<td>0.464</td>
<td>64</td>
</tr>
</tbody>
</table>
Evidence of Catullan Influence in *Angustae Vitae*

Samples classified positively with Catullus out of all samples:

<table>
<thead>
<tr>
<th>More Like Catullus</th>
<th>Positive Class.</th>
<th>Text</th>
<th>Less Like Catullus</th>
<th>Positive Class.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Angustae Vitae</strong></td>
<td>1/1</td>
<td>Ovid <em>Amores</em></td>
<td>2/40</td>
<td></td>
</tr>
<tr>
<td>Propertius <em>Elegies</em></td>
<td>6/40</td>
<td>Horace <em>Epistles</em></td>
<td>3/40</td>
<td></td>
</tr>
<tr>
<td>Tibullus <em>Elegies</em></td>
<td>5/40</td>
<td>Virgil <em>Aeneid</em></td>
<td>2/35</td>
<td></td>
</tr>
</tbody>
</table>
Evidence of Catullan Influence in *Angustae Vitae*

Samples classified positively with Catullus out of all samples after refinement using low-probability features:

<table>
<thead>
<tr>
<th>More Like Catullus</th>
<th>Text</th>
<th>Positive Class.</th>
<th>Less Like Catullus</th>
<th>Text</th>
<th>Positive Class.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>Angustae Vitae</em></td>
<td>1/1</td>
<td></td>
<td>Ovid <em>Amores</em></td>
<td>1/40</td>
</tr>
<tr>
<td></td>
<td>Propertius <em>Elegies</em></td>
<td>5/40</td>
<td></td>
<td>Horace <em>Epistles</em></td>
<td>1/40</td>
</tr>
<tr>
<td></td>
<td>Tibullus <em>Elegies</em></td>
<td>5/40</td>
<td></td>
<td>Virgil <em>Aeneid</em></td>
<td>1/35</td>
</tr>
</tbody>
</table>
Conclusions

Quantitative Intertextuality has the potential to draw out new aspects of stylistic influence and reference.

Functional n-gram feature provided a tool to quantify a notable stylistic similarity in sound between Catullus 1 & 2 and *Angustae Vitae*

Low probability analysis was able to refine results.
Wrapping Up
Further Reading


Code

1-vs-Set Machine on GitHub:
https://github.com/tboult/libSVM-onevset

Data sets:
http://www.metarecognition.com/openset/