Hierarchical Classification of Diatom Images using Predictive Clustering Trees

Ivica Dimitrovski¹, Dragi Kocev², Suzana Loskovska¹, Sašo Džeroski²

¹ Faculty of Electrical Engineering and Information Technologies, Department of Computer Science, Skopje, Macedonia

² Jožef Stefan Institute, Department of Knowledge Technologies, Ljubljana, Slovenia
Outline

- Hierarchical multi-label classification system for diatom image classification
- Contour and feature extraction from images
- Predictive Clustering Trees
- Ensembles: Bagging and random forests
- Experimental Design
- Results and Discussion
Diatom image classification (1)

- Diatoms: large and ecologically important group of unicellular or colonial organisms (algae)
- Variety of uses: water quality monitoring, paleoecology and forensics
Diatom image classification (2)

- 200,000 different diatom species, half of them still undiscovered

- Automatic diatom classification
  - image processing (feature extraction from images)
  - image classification (labels and groups the images)

- Labels can be organized in a hierarchy and an image can be labeled with more than one label

- Predict all different levels in the hierarchy of taxonomic ranks: genus, species, variety, and form

- Goal of the complete system: assist a taxonomist in identifying a wide range of different diatoms
Diatom image classification (3)

- Set of images with their visual descriptors and annotations
- Taxonomic rank with hierarchical structure

<table>
<thead>
<tr>
<th>Image</th>
<th>features/descriptors</th>
<th>taxonomy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Heuristic shape descriptors</td>
<td></td>
</tr>
<tr>
<td></td>
<td>48 24 59 66 37 ...</td>
<td>olyvaceum</td>
</tr>
<tr>
<td></td>
<td>36 25 53 45 15 ...</td>
<td>minutissimum</td>
</tr>
<tr>
<td></td>
<td>35 25 56 52 19</td>
<td>exigua</td>
</tr>
</tbody>
</table>
| ... | ... | ... | ...
Contour extraction from images

● Pre-segmentation of an image
  ● separate the diatom objects from dark or light debris
  ● identify the regions with structured objects
  ● merge nested regions

● Edge-based thresholding for contour extraction
  ● locate the boundary between the objects and the background
  ● produce a binary (black and white) image with the diatom contours

● Contour following
  ● trace the region borders in the binary image
Feature extraction from images

- Simple geometric properties
  - length, width, size and the length-width ratio

- Simple shape descriptors
  - rectangularity, triangularity, compactness, ellipticity, and circularity

- Fourier descriptors
  - 30 coefficients

- SIFT histograms
  - Invariant to changes in illumination, image noise, rotation, scaling, and small changes in viewpoint
Predictive Clustering Trees (PCTs)

- Standard Top-Down Induction of DTs
- Hierarchy of clusters
- Distance measure: minimization of intra-cluster variance
- Instantiation of the variance for different tasks
  - Multiple targets, sequences, hierarchies
**CLUS**

- System where the PCTs framework is implemented (KULeuven & JSI)
- The top-down induction algorithm for PCTs:

```plaintext
procedure PCT(I) returns tree
1: (t*, P*) = BestTest(I)
2: if t* ≠ none then
3:   for each I_k ∈ P* do
4:     tree_k = PCT(I_k)
5:   return node(t*, ∪_k{tree_k})
6: else
7:   return leaf(Prototype(I))
```

```plaintext
procedure BestTest(I)
1: (t*, h*, P*) = (none, 0, Ø)
2: for each possible test t do
3:   P = partition induced by t on I
4:   h = Var(I) - ∑_{I_k ∈ P} |I_k|/|I| Var(I_k)
5:   if (h > h*) ∧ Acceptable(t, P) then
6:     (t*, h*, P*) = (t, h, P)
7:   return (t*, P*)
```

- Selecting the tests: reduction in variance caused by partitioning the instances
PCTs for Hierarchical Multi-Label Classification

- HMLC: an example can be labeled with multiple labels that are organized in a hierarchy

\[
\{ 1, 2, 2.2 \} \quad \rightarrow \quad v_i = [1, 1, 0, 1, 0]
\]
PCTs for Hierarchical Multi-Label Classification

- Variance: average squared distance between each example’s label and the set’s mean label

- Weighted Euclidean distance: an error at the upper levels costs more than an error at the lower levels

\[
Var(S) = \frac{\sum_{i} d(v_i, \bar{v})^2}{|S|}
\]

\[
d(v_1, v_2) = \sqrt{\sum_{i} w(c_i)(v_{1,i} - v_{2,i})^2}
\]
Ensemble methods

- Ensembles are a set of predictive models
  - Unstable base classifiers

- Voting schemes to combine the predictions into a single prediction

- Ensemble learning approaches
  - Modification of the data
    - Bagging
  - Modification of the algorithm
    - Random Forest
Ensemble methods

- Training set
- n bootstrap replicates

n classifiers → n predictions

Test set

L1 → L2 → L3 → L

Ensemble methods

Training set

1

2

3

... n

n bootstrap replicates

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS

CLUS
ADIAC diatom image database

- Three different subset of images:
  - 1099 images classified in 55 different taxa
  - 1020 images classified in 48 different taxa
  - 819 images classified in 37 different taxa

- The diatoms vary in shape and ornamentation
Experimental design – classifier

- Random Forests and Bagging of PCTs for HMLC:
  - Feature Subset Size: 10% of the number of descriptive attributes
  - Number of classifiers: 100 un-pruned trees
  - Combine the predictions output by the base classifiers: probability distribution vote
Results (1)

- Predictive performance of the feature extraction algorithms and their combination

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Descriptors</th>
<th># features</th>
<th>Overall recognition rate [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>55 diatom taxa</td>
</tr>
<tr>
<td><strong>Bagging</strong></td>
<td>Geometric and shape descriptors</td>
<td>9</td>
<td>76.3</td>
</tr>
<tr>
<td></td>
<td>Fourier descriptors</td>
<td>30</td>
<td>86.7</td>
</tr>
<tr>
<td></td>
<td>SIFT histograms</td>
<td>200</td>
<td>88.4</td>
</tr>
<tr>
<td></td>
<td>Geometric and shape desc.+Fourier desc.+SIFT hist.</td>
<td>239</td>
<td>96.2</td>
</tr>
<tr>
<td><strong>Random Forests</strong></td>
<td>Geometric and shape descriptors</td>
<td>9</td>
<td>76.3</td>
</tr>
<tr>
<td></td>
<td>Fourier descriptors</td>
<td>30</td>
<td>86.6</td>
</tr>
<tr>
<td></td>
<td>SIFT histograms</td>
<td>200</td>
<td>88.2</td>
</tr>
<tr>
<td></td>
<td>Geometric and shape desc.+Fourier desc.+SIFT hist.</td>
<td>239</td>
<td>96.2</td>
</tr>
</tbody>
</table>
Results (1)

- Predictive performance of the feature extraction algorithms and their combination

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Descriptors</th>
<th># features</th>
<th>Overall recognition rate [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>55 diatom taxa</td>
</tr>
<tr>
<td>Bagging</td>
<td>Geometric and shape descriptors</td>
<td>9</td>
<td>76.3</td>
</tr>
<tr>
<td></td>
<td>Fourier descriptors</td>
<td>30</td>
<td>86.7</td>
</tr>
<tr>
<td></td>
<td>SIFT histograms</td>
<td>200</td>
<td>88.4</td>
</tr>
<tr>
<td></td>
<td>Geometric and shape desc.+Fourier desc.+SIFT hist.</td>
<td>239</td>
<td>96.2</td>
</tr>
<tr>
<td>Random Forests</td>
<td>Geometric and shape descriptors</td>
<td>9</td>
<td>76.3</td>
</tr>
<tr>
<td></td>
<td>Fourier descriptors</td>
<td>30</td>
<td>86.6</td>
</tr>
<tr>
<td></td>
<td>SIFT histograms</td>
<td>200</td>
<td>88.2</td>
</tr>
<tr>
<td></td>
<td>Geometric and shape desc.+Fourier desc.+SIFT hist.</td>
<td>239</td>
<td>96.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data</th>
<th>Descriptors</th>
<th>Classifier</th>
<th>Evaluation</th>
<th>Recognition Rate [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td># Images</td>
<td># Taxa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1099</td>
<td>55 geometric and shape; Fourier; SIFT</td>
<td>Bagging of predictive clustering trees</td>
<td>10-fold cross-validation</td>
<td>96.2</td>
</tr>
<tr>
<td>1020</td>
<td>48 geometric and shape; Fourier; SIFT</td>
<td>Bagging of predictive clustering trees</td>
<td>10-fold cross-validation</td>
<td>98.1</td>
</tr>
<tr>
<td>1009</td>
<td>48 contour profiling; Legendre polynomials</td>
<td>Decision trees; Neural networks; syntactical classifier</td>
<td>Random separation (50/50) to train and test set</td>
<td>82</td>
</tr>
<tr>
<td>808</td>
<td>38 geometric; shape; Fourier; image moments; ornamentation and morphological</td>
<td>Bagging of Decision Trees</td>
<td>Leave One Out</td>
<td>94.9</td>
</tr>
<tr>
<td>819</td>
<td>37 geometric and shape; Fourier; SIFT</td>
<td>Bagging of predictive clustering trees</td>
<td>10-fold cross-validation</td>
<td>98.8</td>
</tr>
<tr>
<td>781</td>
<td>37 contour; segment; global</td>
<td>nearest -mean classifier</td>
<td>set swapping (complex pseudo cross-validation)</td>
<td>82.9</td>
</tr>
<tr>
<td>781</td>
<td>37 Gabor; Legendre polynomials; ornamentation</td>
<td>Decision trees; Bayesian classifier</td>
<td>Random separation (50/50) to train and test set</td>
<td>88</td>
</tr>
<tr>
<td>781</td>
<td>37 contour; ornamentation</td>
<td>Bagging of Decision Trees</td>
<td>10 times random separation (75/25) train and test</td>
<td>89.6</td>
</tr>
<tr>
<td>781</td>
<td>37 Gabor; Legendre polynomials; ornamentation; contour; global; geometric; shape; Fourier; image moments; morphological</td>
<td>Bagging of Decision Trees</td>
<td>10 times random separation (75/25) train and test</td>
<td>96.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data # Images</th>
<th># Taxa</th>
<th>Descriptors</th>
<th>Classifier</th>
<th>Evaluation</th>
<th>Recognition Rate [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1099</td>
<td>55</td>
<td>geometric and shape; Fourier; SIFT</td>
<td>Bagging of predictive clustering trees</td>
<td>10-fold cross-validation</td>
<td>96.2</td>
</tr>
<tr>
<td>1020</td>
<td>48</td>
<td>geometric and shape; Fourier; SIFT</td>
<td>Bagging of predictive clustering trees</td>
<td>10-fold cross-validation</td>
<td>98.1</td>
</tr>
<tr>
<td>1009</td>
<td>48</td>
<td>contour profiling; Legendre polynomials</td>
<td>Decision trees; Neural networks; syntactical classifier</td>
<td>Random separation (50/50) to train and test set</td>
<td>82</td>
</tr>
<tr>
<td>808</td>
<td>38</td>
<td>geometric; shape; Fourier; image moments; ornamentation and morphological</td>
<td>Bagging of Decision Trees</td>
<td>Leave One Out</td>
<td>94.9</td>
</tr>
<tr>
<td>819</td>
<td>37</td>
<td>geometric and shape; Fourier; SIFT</td>
<td>Bagging of predictive clustering trees</td>
<td>10-fold cross-validation</td>
<td>98.8</td>
</tr>
<tr>
<td>781</td>
<td>37</td>
<td>contour; segment; global</td>
<td>nearest-mean classifier</td>
<td>set swapping (complex pseudo cross-validation)</td>
<td>82.9</td>
</tr>
<tr>
<td>781</td>
<td>37</td>
<td>Gabor; Legendre polynomials; ornamentation</td>
<td>Decision trees; Bayesian classifier</td>
<td>Random separation (50/50) to train and test set</td>
<td>88</td>
</tr>
<tr>
<td>781</td>
<td>37</td>
<td>contour; ornamentation</td>
<td>Bagging of Decision Trees</td>
<td>10 times random separation (75/25) train and test</td>
<td>89.6</td>
</tr>
<tr>
<td>781</td>
<td>37</td>
<td>Gabor; Legendre polynomials; ornamentation; contour; global; geometric; shape; Fourier; image moments; morphological</td>
<td>Bagging of Decision Trees</td>
<td>10 times random separation (75/25) train and test</td>
<td>96.9</td>
</tr>
</tbody>
</table>
Results (2)

- The presented approach has very high predictive performance (ranging from 96.2% to 98.7%)

- Recognition rates of 100% for the majority of taxa

- Lower recognition rates are achieved for taxa that are very similar to each other and difficult to distinguish
  - *Eunotia diatoms* (presented on image), *Fallacia diatoms*

- Our results are better than the ones obtained from human annotators (63.3% recognition rate)
Conclusion

- Novel approach to taxonomic identification of taxa from microscopic images
- Different feature extraction approaches and hierarchical multi-label classification
- Very high predictive performance - the best reported performance on this dataset