Data Mining and Soft Computing

Francisco Herrera

Research Group on Soft Computing and Information Intelligent Systems (SCI²S)
Dept. of Computer Science and A.I.
University of Granada, Spain

Email: herrera@decsai.ugr.es
http://sci2s.ugr.es
http://decsai.ugr.es/~herrera
Summary

1. Introduction to Data Mining and Knowledge Discovery
2. Data Preparation
3. Introduction to Prediction, Classification, Clustering and Association
4. Data Mining - From the Top 10 Algorithms to the New Challenges
5. Introduction to Soft Computing. Focusing our attention in Fuzzy Logic and Evolutionary Computation
6. Soft Computing Techniques in Data Mining: Fuzzy Data Mining and Knowledge Extraction based on Evolutionary Learning
8. Some Advanced Topics I: Classification with Imbalanced Data Sets
9. Some Advanced Topics II: Subgroup Discovery
10. Some advanced Topics III: Data Complexity
Slides used for preparing this talk:

Data Complexity: An Overview and New Challenges

Tin Kam Ho
Bell Labs, Alcatel-Lucent

Joint work with Mitra Basu, Ester Bernado, Martin Law, Albert Orriols
Outline

✓ Motivation

✓ Class ambiguity, dimensionality and boundary complexity

✓ Measures of Geometric Complexity

✓ Domains of Competence of Classifiers

✓ Other studies

✓ Concluding Remarks
Motivation

Automatic Classification

- Classifiers
  - Bayesian classifiers
  - polynomial discriminators
  - nearest-neighbor methods
  - decision trees & forests
  - neural networks
  - genetic algorithms
  - Fuzzy Rule Based Systems
  - support vector machines
  - ensembles and classifier combination

- Why are machines still far from perfect?
- What is still missing in our techniques?
Large Variations in Accuracies of Different Classifiers

<table>
<thead>
<tr>
<th>Application</th>
<th>ZeroR</th>
<th>NN1</th>
<th>NNK</th>
<th>NB</th>
<th>C4.5</th>
<th>PART</th>
<th>SMO</th>
<th>XCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>aud</td>
<td>25.3</td>
<td>76.0</td>
<td>68.4</td>
<td>69.6</td>
<td>79.0</td>
<td><strong>81.2</strong></td>
<td>-</td>
<td>57.7</td>
</tr>
<tr>
<td>aus</td>
<td>55.5</td>
<td>81.9</td>
<td>85.4</td>
<td>77.5</td>
<td>85.2</td>
<td>83.3</td>
<td>84.9</td>
<td><strong>85.7</strong></td>
</tr>
<tr>
<td>bal</td>
<td>45.0</td>
<td>76.2</td>
<td>87.2</td>
<td><strong>90.4</strong></td>
<td>78.5</td>
<td>81.9</td>
<td>-</td>
<td>79.8</td>
</tr>
<tr>
<td>bpa</td>
<td>58.0</td>
<td>63.5</td>
<td>60.6</td>
<td>54.3</td>
<td>65.8</td>
<td>65.8</td>
<td>58.0</td>
<td><strong>68.2</strong></td>
</tr>
<tr>
<td>bps</td>
<td>51.6</td>
<td>83.2</td>
<td>82.8</td>
<td>78.6</td>
<td>80.1</td>
<td>79.0</td>
<td>86.4</td>
<td><strong>83.3</strong></td>
</tr>
<tr>
<td>bre</td>
<td>65.5</td>
<td>96.0</td>
<td><strong>96.7</strong></td>
<td>96.0</td>
<td>95.4</td>
<td>95.3</td>
<td><strong>96.7</strong></td>
<td>96.0</td>
</tr>
<tr>
<td>cmc</td>
<td>42.7</td>
<td>44.4</td>
<td>46.8</td>
<td>50.6</td>
<td>52.1</td>
<td>49.8</td>
<td>-</td>
<td>52.3</td>
</tr>
<tr>
<td>gls</td>
<td>34.6</td>
<td>66.3</td>
<td>66.4</td>
<td>47.6</td>
<td>65.8</td>
<td>69.0</td>
<td>-</td>
<td><strong>72.6</strong></td>
</tr>
<tr>
<td>h-c</td>
<td>54.5</td>
<td>77.4</td>
<td>83.2</td>
<td><strong>83.6</strong></td>
<td>73.6</td>
<td>77.9</td>
<td>-</td>
<td>79.9</td>
</tr>
<tr>
<td>hep</td>
<td>79.3</td>
<td>79.9</td>
<td>80.8</td>
<td>83.2</td>
<td>78.9</td>
<td>80.0</td>
<td><strong>83.9</strong></td>
<td>83.2</td>
</tr>
<tr>
<td>irs</td>
<td>33.3</td>
<td><strong>95.3</strong></td>
<td><strong>95.3</strong></td>
<td>94.7</td>
<td><strong>95.3</strong></td>
<td>95.3</td>
<td>-</td>
<td>94.7</td>
</tr>
<tr>
<td>krk</td>
<td>52.2</td>
<td>89.4</td>
<td>94.9</td>
<td>87.0</td>
<td>98.3</td>
<td>98.4</td>
<td>96.1</td>
<td>98.6</td>
</tr>
<tr>
<td>lab</td>
<td>65.4</td>
<td>81.1</td>
<td>92.1</td>
<td><strong>95.2</strong></td>
<td>73.3</td>
<td>73.9</td>
<td>93.2</td>
<td>75.4</td>
</tr>
<tr>
<td>led</td>
<td>10.5</td>
<td>62.4</td>
<td>75.0</td>
<td>74.9</td>
<td><strong>74.9</strong></td>
<td>75.1</td>
<td>-</td>
<td>74.8</td>
</tr>
<tr>
<td>lym</td>
<td>55.0</td>
<td>83.3</td>
<td>83.6</td>
<td><strong>85.6</strong></td>
<td>77.0</td>
<td>71.5</td>
<td>-</td>
<td>79.0</td>
</tr>
<tr>
<td>mmg</td>
<td>56.0</td>
<td>63.0</td>
<td><strong>65.3</strong></td>
<td>64.7</td>
<td>64.8</td>
<td>61.9</td>
<td>67.0</td>
<td>63.4</td>
</tr>
<tr>
<td>mus</td>
<td>51.8</td>
<td><strong>100.0</strong></td>
<td><strong>100.0</strong></td>
<td>96.4</td>
<td><strong>100.0</strong></td>
<td><strong>100.0</strong></td>
<td><strong>100.0</strong></td>
<td>99.8</td>
</tr>
<tr>
<td>mux</td>
<td>49.9</td>
<td>78.6</td>
<td>99.8</td>
<td>61.9</td>
<td>99.9</td>
<td><strong>100.0</strong></td>
<td>61.6</td>
<td><strong>100.0</strong></td>
</tr>
<tr>
<td>pmi</td>
<td>65.1</td>
<td>70.3</td>
<td>73.9</td>
<td>75.4</td>
<td>73.1</td>
<td>72.6</td>
<td><strong>76.7</strong></td>
<td>76.0</td>
</tr>
<tr>
<td>prt</td>
<td>24.9</td>
<td>34.5</td>
<td>42.5</td>
<td><strong>50.8</strong></td>
<td>41.6</td>
<td>39.8</td>
<td>-</td>
<td>43.7</td>
</tr>
<tr>
<td>seg</td>
<td>14.3</td>
<td><strong>97.4</strong></td>
<td>96.1</td>
<td>80.1</td>
<td>97.2</td>
<td>96.8</td>
<td>-</td>
<td>96.1</td>
</tr>
<tr>
<td>sick</td>
<td>93.8</td>
<td>96.1</td>
<td>96.3</td>
<td>93.3</td>
<td><strong>98.4</strong></td>
<td>97.0</td>
<td>93.8</td>
<td><strong>93.8</strong></td>
</tr>
<tr>
<td>soyb</td>
<td>13.5</td>
<td>89.5</td>
<td>90.3</td>
<td><strong>92.8</strong></td>
<td>91.4</td>
<td>90.3</td>
<td>-</td>
<td>90.3</td>
</tr>
<tr>
<td>tao</td>
<td>49.8</td>
<td><strong>96.1</strong></td>
<td>96.0</td>
<td>80.8</td>
<td>95.1</td>
<td>93.6</td>
<td>-</td>
<td>93.6</td>
</tr>
<tr>
<td>thy</td>
<td>19.5</td>
<td>68.1</td>
<td>65.1</td>
<td>80.6</td>
<td><strong>92.1</strong></td>
<td><strong>92.1</strong></td>
<td>-</td>
<td>92.1</td>
</tr>
<tr>
<td>veh</td>
<td>25.1</td>
<td>69.4</td>
<td>69.7</td>
<td>46.2</td>
<td>73.6</td>
<td>72.6</td>
<td>-</td>
<td>72.6</td>
</tr>
<tr>
<td>vote</td>
<td>61.4</td>
<td>92.4</td>
<td>92.6</td>
<td>90.1</td>
<td>96.3</td>
<td><strong>96.5</strong></td>
<td>-</td>
<td>96.5</td>
</tr>
<tr>
<td>vow</td>
<td>9.1</td>
<td>99.1</td>
<td><strong>96.6</strong></td>
<td>65.3</td>
<td>80.7</td>
<td>78.3</td>
<td>-</td>
<td>78.3</td>
</tr>
<tr>
<td>wne</td>
<td>39.8</td>
<td>95.6</td>
<td>96.8</td>
<td><strong>97.8</strong></td>
<td>94.6</td>
<td>92.9</td>
<td>-</td>
<td>92.9</td>
</tr>
<tr>
<td>zoo</td>
<td>41.7</td>
<td>94.6</td>
<td>92.5</td>
<td><strong>95.4</strong></td>
<td>91.6</td>
<td>92.5</td>
<td>-</td>
<td>92.5</td>
</tr>
<tr>
<td>Avg</td>
<td><strong>44.8</strong></td>
<td><strong>80.0</strong></td>
<td><strong>82.4</strong></td>
<td>78.0</td>
<td>82.1</td>
<td>81.8</td>
<td><strong>84.1</strong></td>
<td><strong>81.7</strong></td>
</tr>
</tbody>
</table>
Many classifiers are in close rivalry with each other. Why?

- Do they represent the limit of our technology?
- What do the new classifiers add to the methodology?
- Is there still value in the older methods?
- Have they used up all information contained in a data set?

When I face a new recognition task ...

- How much can automatic classifiers do?
- How should I choose a classifier?
- Can I make the problem easier for a specific classifier?
Complexity Measures

Sources of Difficulty in Classification

- Class ambiguity
- Sample size and dimensionality
- Boundary complexity

We need metrics for analyzing problems features and the limits of every learning model.

Limits of Current Learning Algorithms
Some Advanced Topics III: Data Complexity

Outline

✓ Motivation

✓ Class ambiguity, dimensionality and boundary complexity

✓ Measures of Geometric Complexity

✓ Domains of Competence of Classifiers

✓ Other studies

✓ Concluding Remarks
Class Ambiguity

- Is the concept intrinsically ambiguous?
- Are the classes well defined?

- What information do the features carry?
- Are the features sufficient for discrimination?

Bayes error
Problem may appear deceptively simple or complex with small samples.
Boundary Complexity

- Kolmogorov complexity
- Length can be exponential in dimensionality
- A trivial description is to list all points & class labels
- Is there a shorter description?
Classification Boundaries As Decided by Different Classifiers

Training samples for a 2D classification problem
Classification Boundaries Inferred by Different Classifiers

- XCS: a genetic algorithm
- Nearest neighbor classifier
- Linear classifier
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%
Some Advanced Topics III: Data Complexity

Outline

✓ Motivation
✓ Class ambiguity, dimensionality and boundary complexity
✓ Measures of Geometric Complexity
✓ Domains of Competence of Classifiers
✓ Other studies
✓ Concluding Remarks
Measures of Geometrical Complexity of Classification Problems

The approach: develop mathematical language and algorithmic tools for studying

- Characteristics of geometry & topology of high-dim data
- How they change with feature transformations, noise conditions, and sampling strategies
- How they interact with classifier geometry

Focus on descriptors computable from real data and relevant to classifier geometry
Geometry of Datasets and Classifiers

- **Data sets:**
  - length of class boundary
  - fragmentation of classes / existence of subclasses
  - global or local linear separability
  - convexity and smoothness of boundaries
  - intrinsic / extrinsic dimensionality
  - stability of these characteristics as sampling rate changes

- **Classifier models:**
  - polygons, hyper-spheres, Gaussian kernels, axis-parallel hyper-planes, piece-wise linear surfaces, polynomial surfaces, their unions or intersections, ...
# Measures of Geometric Complexity

## Degree of Linear Separability
- Find separating hyper-plane by linear programming
- Error counts and distances to plane measure separability

## Fisher’s Discriminant Ratio
- Classical measure of class separability
- Maximize over all features to find the most discriminating

\[
 f = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}
\]

## Length of Class Boundary
- Compute minimum spanning tree
- Count class-crossing edges

## Shapes of Class Manifolds
- Cover same-class pts with maximal balls
- Ball counts describe shape of class manifold
## 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>
Example

Method Ishibuchi FH-GGBML, 2005, IEEE TSMC

<table>
<thead>
<tr>
<th>Measure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>F2</td>
<td>volume of overlap region</td>
</tr>
<tr>
<td>L1</td>
<td>minimized sum of error distance by linear programming</td>
</tr>
<tr>
<td>L2</td>
<td>error rate of linear classifier by Linear Programming</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>non-linearity of 1NN classifier</td>
</tr>
<tr>
<td>T2</td>
<td>average number of points per dimension</td>
</tr>
</tbody>
</table>

Table 1: Complexity metrics used in this study

Figure 2: Accuracy in Train/Test for FH-GBML sorted by train accuracy

255 data sets with 2-variables, nonseparable Benchmarking data from UC-Irvine archive
Method Ishibuchi FH-GGBML

Figure 3: Accuracy in Train/Test sorted by F2

Figure 4: Accuracy in Train/Test sorted by N2

Figure 5: Accuracy in Train/Test sorted by N3

Figure 6: Accuracy in Train/Test sorted by N4
Method Ishibuchi FH-GGBML

Figure 7: Accuracy in Train/Test sorted by L1

Figure 8: Accuracy in Train/Test sorted by L2

Figure 9: Accuracy in Train/Test sorted by T2
Some Advanced Topics III: Data Complexity

Outline

✓ Motivation

✓ Class ambiguity, dimensionality and boundary complexity

✓ Measures of Geometric Complexity

✓ Domains of Competence of Classifiers

✓ Other studies

✓ Concluding Remarks
Domains of Competence of Classifiers

- Given a classification problem, determine which classifier is the best for it.

![Diagram showing different complexity measures with classifiers XCS, NN, and LC overlapping and labeled Fuzzy Systems]
Domains of Competence of Classifiers

Method Ishibuchi  FH-GGBML

Some interesting intervals

<table>
<thead>
<tr>
<th>Interval</th>
<th>FH-GBML Behaviour</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_2 &lt; 0.23$</td>
<td>good behaviour</td>
</tr>
<tr>
<td>$L_1 &lt; 0.1585$</td>
<td>good behaviour</td>
</tr>
<tr>
<td>$F_2 = 1$</td>
<td>good behaviour</td>
</tr>
<tr>
<td>$0.04 &lt; L_2 &lt; 0.1$</td>
<td>good behaviour</td>
</tr>
<tr>
<td>$N_3 = 0$</td>
<td>bad behaviour</td>
</tr>
<tr>
<td>$N_4 = 0$</td>
<td>bad behaviour</td>
</tr>
<tr>
<td>$T_2 &lt; 7$</td>
<td>bad behaviour</td>
</tr>
</tbody>
</table>

Table 2: Significant intervals
## Domains of Competence of Classifiers

### Method Ishibuchi FH-GGBML

**Rules with a metric**

<table>
<thead>
<tr>
<th>Id.</th>
<th>Rule</th>
<th>Support</th>
<th>Avg. % Train St.Dev</th>
<th>Train Diff.</th>
<th>Avg. % Test St.Dev</th>
<th>Test Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1+</td>
<td>If $N2[X] &lt; 0.23$ then good behaviour</td>
<td>32.549%</td>
<td>99.10000% 1.56873</td>
<td>6.8880%</td>
<td>96.40400% 3.73928</td>
<td>12.6190%</td>
</tr>
<tr>
<td>R2+</td>
<td>If $L1[X] &lt; 0.1585$ then good behaviour</td>
<td>16.471%</td>
<td>98.79382% 1.88762</td>
<td>6.5810%</td>
<td>96.63110% 6.92474</td>
<td>12.8459%</td>
</tr>
<tr>
<td>R3+</td>
<td>If $F2[X] = 1$ then good behaviour</td>
<td>19.216%</td>
<td>95.99478% 4.08713</td>
<td>3.7820%</td>
<td>91.47715% 5.74098</td>
<td>7.6919%</td>
</tr>
<tr>
<td>R4+</td>
<td>If $0.04 &lt; L2[X] &lt; 0.1$ then good behaviour</td>
<td>19.608%</td>
<td>97.07823% 2.46866</td>
<td>4.8654%</td>
<td>91.73752% 6.76988</td>
<td>7.9523%</td>
</tr>
<tr>
<td>R1-</td>
<td>If $N3[X] = 0$ then bad behaviour</td>
<td>18.039%</td>
<td>90.17976% 28.26869</td>
<td>-2.03303%</td>
<td>78.79163% 30.81635</td>
<td>-4.99360%</td>
</tr>
<tr>
<td>R2-</td>
<td>If $N4[X] = 0$ then bad behaviour</td>
<td>27.059%</td>
<td>88.73440% 30.12516</td>
<td>-3.47839%</td>
<td>77.14338% 31.48844</td>
<td>-6.64185%</td>
</tr>
<tr>
<td>R3-</td>
<td>If $T2[X] &lt; 7$ then bad behaviour</td>
<td>30.588%</td>
<td>86.47399% 29.72216</td>
<td>-5.73880%</td>
<td>69.42453% 28.89741</td>
<td>-14.36070%</td>
</tr>
</tbody>
</table>

Table 3: Rules with one metric obtained from the intervals
Domains of Competence of Classifiers

**Method Ishibuchi FH-GGBML**

**Rules with a metric**

<table>
<thead>
<tr>
<th>Id.</th>
<th>Rule</th>
<th>Support</th>
<th>Avg. % Train St.Dev</th>
<th>Train Diff.</th>
<th>Avg. % Test St.Dev</th>
<th>Test Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>R5+</td>
<td>If L1[X] &lt; 0.1585 and not T2[X] &lt; 7 then good behaviour</td>
<td>10.196%</td>
<td>98.72043% 1.72081</td>
<td>6.5076%</td>
<td>97.29695% 2.3808</td>
<td>13.5117%</td>
</tr>
<tr>
<td>R6+</td>
<td>If N2[X] &lt; 0.1585 and not N3[X] = 0 then good behaviour</td>
<td>22.353%</td>
<td>98.68990% 1.74822</td>
<td>6.4771%</td>
<td>95.46808% 3.88134</td>
<td>11.6829%</td>
</tr>
<tr>
<td>R7+</td>
<td>If 0.04 &lt; L2[X] &lt; 0.1 and not T2[X] &lt; 7 then good behaviour</td>
<td>14.902%</td>
<td>96.88916% 2.22073</td>
<td>4.6764%</td>
<td>93.02681% 4.67047</td>
<td>9.2416%</td>
</tr>
<tr>
<td>R4-</td>
<td>If N3[X] = 0 and not L1[X] &lt; 0.19 then bad behaviour</td>
<td>12.941%</td>
<td>86.45058% 32.71477</td>
<td>-5.76221%</td>
<td>71.02749% 33.35019</td>
<td>-12.75774%</td>
</tr>
<tr>
<td>R5-</td>
<td>If N3[X] = 0 and not N2[X] &lt; 0.23 then bad behaviour</td>
<td>7.843%</td>
<td>77.64346% 39.94092</td>
<td>-14.56933%</td>
<td>53.22919% 32.01059</td>
<td>-30.55604%</td>
</tr>
<tr>
<td>R6-</td>
<td>If N4[X] = 0 and not L1[X] &lt; 0.19 then bad behaviour</td>
<td>20.000%</td>
<td>84.82022% 34.26575</td>
<td>-7.39257%</td>
<td>69.74147% 33.6301</td>
<td>-14.04376%</td>
</tr>
<tr>
<td>R7-</td>
<td>If N4[X] = 0 and not N2[X] &lt; 0.23 then bad behaviour</td>
<td>14.510%</td>
<td>79.00123% 38.78489</td>
<td>-13.21156%</td>
<td>59.12644% 33.87836</td>
<td>-24.65879%</td>
</tr>
</tbody>
</table>

Table 4: Rules with two metrics obtained from the intervals
Domains of Competence of Classifiers

Method Ishibuchi  FH-GGBML
Combination of Rules

<table>
<thead>
<tr>
<th>Id.</th>
<th>Rule</th>
<th>Support</th>
<th>Avg. % Train St.Dev</th>
<th>Train Diff.</th>
<th>Avg. % Test St.Dev</th>
<th>Test Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDP</td>
<td>If R1+ or R2+ or R3+ or R4+ or R5+ or R6+ or R7+ then good behaviour</td>
<td>50.196%</td>
<td>97.87275% 3.24086</td>
<td>5.65996%</td>
<td>94.10161% 6.21307</td>
<td>10.31638%</td>
</tr>
<tr>
<td>RDN</td>
<td>If R1- or R2- or R3- or R4- or R5- or R6- or R7- then bad behaviour</td>
<td>41.176%</td>
<td>89.77980% 26.23892</td>
<td>-2.43299%</td>
<td>76.29024% 27.76105</td>
<td>-7.49499%</td>
</tr>
</tbody>
</table>

Table 6: Disjunction Rules from all rules

<table>
<thead>
<tr>
<th>Id.</th>
<th>Rule</th>
<th>Support</th>
<th>Avg. % Train St.Dev</th>
<th>Train Diff.</th>
<th>Avg. % Test St.Dev</th>
<th>Test Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDP \land RDN</td>
<td>If RDP and RDN then good behaviour</td>
<td>18.824%</td>
<td>99.41149% 1.83755</td>
<td>7.19870%</td>
<td>95.18885% 7.09706</td>
<td>11.40303%</td>
</tr>
<tr>
<td>RDP \lor \lnot RDN</td>
<td>If RDP and not RDN then good behaviour</td>
<td>31.373%</td>
<td>96.94950% 3.546016</td>
<td>4.73671%</td>
<td>93.44961% 5.56263</td>
<td>9.66438%</td>
</tr>
<tr>
<td>RDN \lor \lnot RDP</td>
<td>If RDN and not RDP then bad behaviour</td>
<td>22.353%</td>
<td>81.66890% 33.60499</td>
<td>-10.54389%</td>
<td>60.37611% 28.72427</td>
<td>-23.40912%</td>
</tr>
</tbody>
</table>

Table 7: Intersections of the disjunction rules
Domains of Competence of Classifiers

Method Ishibuchi  FH-GGBML
Combination of Rules – Behaviour Caracterization

- RDP - good behaviour 50.196%
- RDN^~RDP - bad behaviour 22.353%
- not characterized 27.451%
### Domains of Competence of Classifiers

**Method Ishibuchi FH-GGBML**

**Combination of Rules – Behaviour Characterization**

<table>
<thead>
<tr>
<th>Id.</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDP</td>
<td>If ((N_2[X] &lt; 0.23) ) or ((L_1[X] &lt; 0.1585) ) or ((F_2[X] = 1) ) or ((0.04 &lt; L_2[X] &lt; 0.1) ) or ((L_1[X] &lt; 0.1585 \text{ and not } T_2[X] &lt; 7) ) or ((N_2[X] &lt; 0.1585 \text{ and not } N_3[X] = 0) ) or ((0.04 &lt; L_2[X] &lt; 0.1 \text{ and not } T_2[X] &lt; 7) ) then good behaviour</td>
</tr>
<tr>
<td>RDN(\land)RDP</td>
<td>If ([((N_3[X] = 0) \text{ or } (N_4[X] = 0) \text{ or } (T_2[X] &lt; 7) \text{ or } (N_3[X] = 0 \text{ and not } L_1[X] &lt; 0.19) \text{ or } (N_3[X] = 0 \text{ and not } N_2[X] &lt; 0.23) \text{ or } (N_4[X] = 0 \text{ and not } L_1[X] &lt; 0.19) \text{ or } (N_4[X] = 0 \text{ and not } N_2[X] &lt; 0.23)]) \text{ and not } [(N_2[X] &lt; 0.23) \text{ or } (L_1[X] &lt; 0.1585) \text{ or } (F_2[X] = 1) \text{ or } (0.04 &lt; L_2[X] &lt; 0.1) \text{ or } (L_1[X] &lt; 0.1585 \text{ and not } T_2[X] &lt; 7) \text{ or } (N_2[X] &lt; 0.1585 \text{ and not } N_3[X] = 0) \text{ or } (0.04 &lt; L_2[X] &lt; 0.1 \text{ and not } T_2[X] &lt; 7)] \text{ then bad behaviour}</td>
</tr>
</tbody>
</table>

Table 8: RDP and RDN\(\land\)RDP rules
Domains of Competence of Classifiers

Comparison of classifiers with a measure

Best Classifier for Benchmarking Data
Domains of Competence of Classifiers

Best Classifier Being 
\textbf{nn,lp,odt vs an ensemble technique}

\textbf{Boundary-NonLinNN}

\textbf{IntralInter-Pretop}

\textbf{MaxEff-VolumeOverlap}

ensemble
+ nn,lp,odt
Outline

✓ Motivation

✓ Class ambiguity, dimensionality and boundary complexity

✓ Measures of Geometric Complexity

✓ Domains of Competence of Classifiers

✓ Other studies

✓ Concluding Remarks
Complexity and Sample Sparsity

Sparse Sample & complex geometry cause ill-posedness

Careful statistical procedures are needed to infer complexity of the data population from those of the training samples

Complexity estimation requires further hypotheses on data geometry and sampling processes
Complexity and Data Dimensionality: Class Separability after Dimensionality Reduction

- Feature selection 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
Extensions of the Study on Data Complexity

Multi-Class Measures

Global vs. Local Properties

Intrinsic Ambiguity & Mislabeling

Task Trajectory with Changing Sampling & Noise Conditions

Diagram of network structures and data points illustrating intrinsic ambiguity and task trajectory with changing conditions.
Extension to Multiple Classes

- Fisher’s discriminant score $\Rightarrow$ Multiple discriminant scores

\[
 f = \frac{(\mu_1 - \mu_2)^2}{(\sigma_1^2 + \sigma_2^2)} \quad \Rightarrow \quad f = \frac{\sum_{i=1, j=1, i \neq j}^{C} p_i p_j (\mu_i - \mu_j)^2}{\sum_{i=1}^{C} p_i \sigma_i^2}
\]

- Boundary point in a MST: a point is a boundary point as long as it is next to a point from other classes in the MST
Comparing Global vs. Local Properties

- Boundaries can be simple locally but complex globally
  - These types of problems are relatively simple, but are characterized as complex by the measures

- Solution: complexity measure at different scales
  - This can be combined with different error levels

- Let $N_{i,k}$ be the $k$ neighbors of the $i$-th point defined by, say, Euclidean distance. The complexity measure for data set $D$, error level $\varepsilon$, evaluated at scale $k$ is

$$\bar{f}(D, \varepsilon, k) = \frac{1}{n} \sum_{i=1}^{n} f(N_{i,k}, \varepsilon)$$
Effects of Intrinsic Ambiguity

- The complexity measures can be severely affected when there exists intrinsic class ambiguity (or data mislabeling)
  - Example: FeatureOverlap (in 1D only)

- Cannot distinguish between intrinsic ambiguity or complex class decision boundary
Tackling Intrinsic Ambiguity

- Compute the complexity measure at different error levels
  - $f(D)$: a complexity measure on the data set $D$
  - $D^*$: a “perturbed” version of $D$, so that some points are relabeled
  - $h(D, D^*)$: a distance measure between $D$ and $D^*$ (error level)
- The new complexity measure is defined as a curve:
  \[
g(D, \epsilon) = \min_{D^* : h(D, D^*) \leq \epsilon} f(D^*)\]
- The curve can be summarized by, say, area under curve

- Minimization by greedy procedures
  - Discard erroneous points that decrease complexity by the most
Outline

✓ Motivation

✓ Class ambiguity, dimensionality and boundary complexity

✓ Measures of Geometric Complexity

✓ Domains of Competence of Classifiers

✓ Other studies

✓ Concluding Remarks
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.
- Difficulties are class ambiguity, geometric complexity, & sample sparsity.
- Measures for geometric complexity are useful to characterize classifier domains of competence.
Better understanding of how data and classifiers interact can guide practice.

Further progress in statistical and machine learning will need systematic, scientific evaluation of the algorithms with problems that are difficult for different reasons.
Summary

1. Introduction to Data Mining and Knowledge Discovery
2. Data Preparation
3. Introduction to Prediction, Classification, Clustering and Association
4. Data Mining - From the Top 10 Algorithms to the New Challenges
5. Introduction to Soft Computing. Focusing our attention in Fuzzy Logic and Evolutionary Computation
6. Soft Computing Techniques in Data Mining: Fuzzy Data Mining and Knowledge Extraction based on Evolutionary Learning
8. Some Advanced Topics I: Classification with Imbalanced Data Sets
9. Some Advanced Topics II: Subgroup Discovery
10. Some advanced Topics III: Data Complexity