A Simple Methodology for Soft Cost-sensitive Classification

Te-Kang Jan
tekang@iis.sinica.edu.tw

Da-Wei Wang
wdw@iis.sinica.edu.tw

Chi-Hung Lin
linch@ym.edu.tw

Hsuan-Tien Lin
htlin@csie.ntu.edu.tw

Proposed Method:
Soft Cost-sensitive Classification

- Apply weighted sum approach for multicriteria optimization: a popular and simple approach

\[
\min \, g \sum_{m=1}^{M} \alpha_m E_m(g)
\]

- Benefits:
  - Transform multicriteria to a single criterion: easier to solve
  - Allow reusing of existing cost-sensitive classification algorithms: immediate improvement

More specifically, we solve

\[
\min \, \mathcal{E} (1 - \alpha) \left( \mathcal{C}[g(x)] \right) + \alpha \left( \mathcal{E}_w[g(x)] \right)
\]

— cost-sensitive classification with linear combination of original cost vector and classification cost vector:

- \( \alpha = 0 \): Hard cost-sensitive — minimize cost
- \( \alpha = 1 \): Non cost-sensitive — minimize error rates
- \( 0 < \alpha < 1 \): Soft cost-sensitive — minimize both

Experiment Result: Benchmark Datasets
- 22 UCI datasets with benchmark artificial cost
- Compare to a leading hard cost-sensitive classifier (OSR)

Table: Error Measure = Treatment Cost

<table>
<thead>
<tr>
<th>Classify to Actual</th>
<th>Bacteria 1</th>
<th>Bacteria 2</th>
<th>Bacteria 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteria 1</td>
<td>0</td>
<td>30</td>
<td>10000</td>
</tr>
<tr>
<td>Bacteria 2</td>
<td>100</td>
<td>0</td>
<td>6000</td>
</tr>
<tr>
<td>Bacteria 3</td>
<td>7000</td>
<td>5000</td>
<td>0</td>
</tr>
</tbody>
</table>

Goal for This Application
- Low cost: To save patients realistically, find a classifier \( g \) with small

\[ E_c(g) = \mathcal{E} \, c[g(x)], \]

for future unseen case (patient \( x \), cost \( c \))

- Low error rates: To make doctors happily use the system, need small

\[ E(g) = \mathcal{E} \, \hat{e}[y|x], \]

for future unseen case (patient \( x \), bacteria \( y \))

- Application goal: simultaneously

\[ \min_g E(g) = [E_c(g), E(g)] \]

subject to all feasible \( g \)

Motivation
- Trade-off between cost and error rates:

Green area: low error rates but the high cost

Red area: low cost but the high error rates

Blue area: our goal, low cost and low error rates

Multicriteria Optimization
Key idea: view the task as

\[ \min_g E(g) = [E_1(g), E_2(g), \ldots, E_M(g)] \]

where \( M \) is the number of criteria.

Summary
- Explore the trade-off between error rates and cost by taking both criteria into account
- Provide a new insight for cost-sensitive classification — feeding the exact cost information may not be the best
- Achieve the best performance for the real-world medical application