Cost-sensitive Multiclass Classification Using One-versus-one Comparisons

Hsuan-Tien Lin
Assistant Professor
Dept. of Computer Science and Information Engineering
National Taiwan University

Talk at ICISE2, 06/24/2012

Which Digit Did You Write?

1  2  3  4
one (1)  two (2)  three (3)  four (4)

- a classification problem
  —grouping “pictures” into different “categories”

How can machines learn to classify?
Learning from Data \cite{abu-mostafa2012learning}

\[ \text{Truth } f(x) + \text{noise } e(x) \]

\[ \text{examples (picture } x_n, \text{ category } y_n) \]

\[ \text{learning algorithm } \]

\[ \text{good decision function } g(x) \approx f(x) \]

\[ \text{learning model } \{ g_\alpha(x) \} \]

\text{challenge: see only } \{(x_n, y_n)\} \text{ without knowing } f(x) \text{ or } e(x) \rightarrow \text{generalize to unseen } (x, y) \text{ w.r.t. } f(x)\]
Cost-Sensitive Classification

Mis-prediction Costs \( g(x) \approx f(x) ? \)

2

ZIP code recognition:
1: wrong; 2: right; 3: wrong; 4: wrong

check value recognition:
1: one-dollar mistake; 2: no mistake;
3: one-dollar mistake; 4: two-dollar mistake

evaluation by formation similarity:
1: not very similar; 2: very similar;
3: somewhat similar; 4: a silly prediction

different applications evaluate mis-predictions differently
ZIP Code Recognition

2
?

1: wrong; 2: right; 3: wrong; 4: right

- **regular** classification problem: only right or wrong
- wrong cost: 1; right cost: 0
- prediction error of \( g \) on some \((x, y)\):

\[
\text{classification cost} = \mathbb{1}[y \neq g(x)]
\]

regular classification: **well-studied**, many good algorithms
Check Value Recognition

2

1: one-dollar mistake; 2: no mistake;
3: one-dollar mistake; 4: two-dollar mistake

- **cost-sensitive** classification problem: different costs for different mis-predictions
- e.g. prediction error of $g$ on some $(x, y)$:

\[\text{absolute cost} = |y - g(x)|\]

cost-sensitive classification: **new**, need more research
What is the Status of the Patient?

H1N1-infected  cold-infected  healthy

Are all mis-prediction costs equal?

another classification problem
—grouping “patients” into different “status”
Cost-Sensitive Classification

Patient Status Prediction

error measure = society cost

\[
C = \begin{array}{ccc}
\text{actual} & \text{predicted} \\
\text{H1N1} & 0 & 1000 & 100000 \\
\text{cold} & 100 & 0 & 3000 \\
\text{healthy} & 100 & 30 & 0 \\
\end{array}
\]

- H1N1 mis-predicted as healthy: **very high cost**
- cold mis-predicted as healthy: **high cost**
- cold correctly predicted as cold: **no cost**

human doctors consider costs of decision; can computer-aided diagnosis do the same?
### Cost Matrix $C$

**regular classification**

$C = \text{classification cost } C_c: \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$

**cost-sensitive classification**

$C = \text{anything other than } C_c: \begin{pmatrix} 0 & 1 & 4 & 5 \\ 1 & 0 & 1 & 3 \\ 3 & 1 & 0 & 2 \\ 5 & 4 & 1 & 0 \end{pmatrix}$

**regular classification:** special case of cost-sensitive classification
Cost-sensitive Classification Setup

**Given**

Given $N$ examples, each (input $x_n$, label $y_n$) $\in \mathcal{X} \times \{1, 2, \ldots, K\} \times \mathbb{R}^K$; cost matrix $C$

- $K = 2$: binary; $K > 2$: **multiclass**
- will assume $C(y, y) = \min_{1 \leq k \leq K} C(y, k)$

**Goal**

a classifier $g(x)$ that pays a small cost $C(y, g(x))$ on future **unseen** example $(x, y)$

cost-sensitive classification: **more realistic than regular one**
Our Contribution

<table>
<thead>
<tr>
<th></th>
<th>binary</th>
<th>multiclass</th>
</tr>
</thead>
<tbody>
<tr>
<td>regular</td>
<td>well-studied</td>
<td>well-studied</td>
</tr>
<tr>
<td>cost-sensitive</td>
<td>known (Zadrozny, 2003)</td>
<td>ongoing (our work, among others)</td>
</tr>
</tbody>
</table>

A theoretical and algorithmic study of cost-sensitive classification, which...

- introduces a methodology for extending regular classification algorithms to cost-sensitive ones with any cost
- provides strong theoretical support for the methodology
- leads to some promising algorithms with superior experimental results

will describe the methodology and a concrete algorithm
Central Idea: Reduction

complex cost-sensitive problems

(iPod)

(adapter)

(cassette player)

simpler regular classification problems with well-known results on models, algorithms, and theories

If I have seen further it is by standing on the shoulders of Giants—I. Newton
Cost-Sensitive Binary Classification (1/2)

medical profile $x$

? 

medical profile $x_1$

H1N1 (1)

predicting H1N1 as NOH1N1: serious consequences to public health

predicting NOH1N1 as H1N1: not good, but less serious

cost-sensitive $C$: 

\[
\begin{pmatrix}
0 & 1000 \\
1 & 0
\end{pmatrix}
\]

regular $C_c$: 

\[
\begin{pmatrix}
0 & 1 \\
1 & 0
\end{pmatrix}
\]

how to change the entry from 1 to 1000?
copy each case labeled **H1N1** 1000 times

<table>
<thead>
<tr>
<th>Original problem</th>
<th>Equivalent problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluate w/ $\begin{pmatrix} 0 &amp; 1000 \ 1 &amp; 0 \end{pmatrix}$</td>
<td>Evaluate w/ $\begin{pmatrix} 0 &amp; 1 \ 1 &amp; 0 \end{pmatrix}$</td>
</tr>
<tr>
<td>$(x_1, H1N1)$</td>
<td>$(x_1, H1N1), \ldots, (x_1, H1N1)$</td>
</tr>
<tr>
<td>$(x_2, \text{NOH1N1})$</td>
<td>$(x_2, \text{NOH1N1})$</td>
</tr>
<tr>
<td>$(x_3, \text{NOH1N1})$</td>
<td>$(x_3, \text{NOH1N1})$</td>
</tr>
<tr>
<td>$(x_4, \text{NOH1N1})$</td>
<td>$(x_4, \text{NOH1N1})$</td>
</tr>
<tr>
<td>$(x_5, H1N1)$</td>
<td>$(x_5, H1N1), \ldots, (x_5, H1N1)$</td>
</tr>
</tbody>
</table>

Mathematically:

$$\begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1000 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$
Key Idea: Cost Transformation

\[
\begin{pmatrix}
0 & 1000 \\
1 & 0
\end{pmatrix}
\begin{pmatrix}
0 & 1 \\
1 & 0
\end{pmatrix}
\begin{pmatrix}
0 & 1000 \\
0 & 1
\end{pmatrix}
\begin{pmatrix}
0 & 1 \\
1 & 0
\end{pmatrix}
\]

\[
\begin{pmatrix}
0 & 1 & 1 & 1 \\
3 & 2 & 3 & 4 \\
1 & 1 & 0 & 1 \\
1 & 1 & 1 & 0
\end{pmatrix}
\begin{pmatrix}
0 & 1 & 1 & 1 \\
1 & 2 & 1 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
0 & 1 & 1 & 1 \\
1 & 0 & 1 & 1 \\
1 & 1 & 0 & 1 \\
1 & 1 & 1 & 0
\end{pmatrix}
\]

- **split** the cost-sensitive example:
  - \((x, 2)\)
  - \(\implies\) a mixture of regular examples \\{\((x, 1), (x, 2), (x, 2), (x, 3)\)\}
  - or a weighted mixture \\{\((x, 1, 1), (x, 2, 2), (x, 3, 1)\)\}

**why split?**
Cost Equivalence by Splitting

\[
\begin{pmatrix}
0 & 1 & 1 & 1 \\
3 & 2 & 3 & 4 \\
1 & 1 & 0 & 1 \\
1 & 1 & 1 & 0 \\
\end{pmatrix}
= 
\begin{pmatrix}
1 & 0 & 0 & 0 \\
1 & 2 & 1 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
\end{pmatrix} \cdot 
\begin{pmatrix}
0 & 1 & 1 & 1 \\
1 & 0 & 1 & 1 \\
1 & 1 & 0 & 1 \\
1 & 1 & 1 & 0 \\
\end{pmatrix}
\]

- \((x, 2)\) implies a weighted mixture \{\((x, 1, 1)\), \((x, 2, 2)\), \((x, 3, 1)\)\}
- **cost equivalence**: for any classifier \(g\),

\[
C(y, g(x)) = \sum_{\ell=1}^{K} Q(y, \ell) C_c(\ell, g(x))
\]

\[
\min_g \text{ expected LHS (cost-sensitive)}
= \min_g \text{ expected RHS (regular when } Q(y, \ell) \geq 0)\
\]
split each training example \((x_n, y_n)\) to a weighted mixture \(\{(x_n, \ell, Q(y_n, \ell))\}^{\ell=1}_{\ell=1} K\)

apply regular classification algorithm on the weighted mixtures \(\bigcup_{n=1}^{N} \{(x_n, \ell, Q(y_n, \ell))\}^{\ell=1}_{\ell=1} K\)

by cost equivalence,

- good \(g\) for new regular classification problem
- good \(g\) for original cost-sensitive classification problem

regular classification: needs \(Q(y_n, \ell) \geq 0\)

but what if \(Q(y_n, \ell)\) negative?
## Similar Cost Vectors

\[
\begin{bmatrix}
1 & 0 & 1 & 2 \\
3 & 2 & 3 & 4
\end{bmatrix}
= \begin{bmatrix}
1/3 & 4/3 & 1/3 & -2/3 \\
1 & 2 & 1 & 0
\end{bmatrix}
\cdot
\begin{bmatrix}
0 & 1 & 1 & 1 \\
1 & 0 & 1 & 1 \\
1 & 1 & 0 & 1 \\
1 & 1 & 1 & 0
\end{bmatrix}
\]

- **negative** \(Q(y, \ell)\): cannot split
- but \(\hat{c} = (1, 0, 1, 2)\) is **similar** to \(c = (3, 2, 3, 4)\):
  - for any classifier \(g\),
    \[
    \hat{c}[g(x)] + \text{constant} = c[g(x)]
    \]
- constant can be dropped during minimization

*shifting cost matrix by constant rows does not affect minimization*
Cost Transformation Methodology

Cost Transformation Methodology: Revised

\[
\begin{pmatrix}
0 & 1 & 1 & 1 \\
1 & 0 & 1 & 2 \\
1 & 1 & 0 & 1 \\
1 & 1 & 1 & 0 \\
\end{pmatrix}
\]
\[
\begin{pmatrix}
0 & 0 & 0 & 0 \\
2 & 2 & 2 & 2 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
\end{pmatrix}
\]
\[
\begin{pmatrix}
1 & 0 & 0 & 0 \\
1 & 2 & 1 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
\end{pmatrix}
\cdot
\begin{pmatrix}
0 & 1 & 1 & 1 \\
1 & 0 & 1 & 1 \\
1 & 1 & 0 & 1 \\
1 & 1 & 1 & 0 \\
\end{pmatrix}
\]

1. Shift each row of original cost to a similar and "splittable" \( C(y,:) \), i.e., with \( Q(y_n, \ell) \geq 0 \)

2. Split \((x_n, y_n)\) to weighted mixture \( \{(x_n, \ell, Q(y_n, \ell))\}^K_{\ell=1} \)

3. Apply regular classification algorithm on the weighted mixtures \( \bigcup_{n=1}^{N} \{(x_n, \ell, Q(y_n, \ell))\}^K_{\ell=1} \)

---

Good \( g \) for new regular classification problem

Good \( g \) for cost-sensitive classification problem

H.-T. Lin (NTU CSIE)  Cost-sensitive One-versus-one  06/24/2012  18 / 25
a single example \(\{(x, 2)\}\) —**certain** that the desired label is 2

a mixture \(\{(x, 1, 1), (x, 2, 2), (x, 3, 1)\}\) sharing the same \(x\) —**uncertainty** in the desired label (25%: 1, 50%: 2, 25%: 3)

over-shifting adds unnecessary mixture uncertainty:

\[
\begin{pmatrix}
3 & 2 & 3 & 4 \\
33 & 32 & 33 & 34
\end{pmatrix}
= \begin{pmatrix}
1 & 2 & 1 & 0 \\
11 & 12 & 11 & 10
\end{pmatrix}
\begin{pmatrix}
0 & 1 & 1 & 1 \\
1 & 0 & 1 & 1 \\
1 & 1 & 0 & 1 \\
1 & 1 & 1 & 0
\end{pmatrix}
\]

should choose a similar and splittable \(c\) with **minimum mixture uncertainty**
Cost Transformation Methodology: Final

1. Shift original cost to a similar and splittable $C$ with minimum "mixture uncertainty."

2. Split $(x_n, y_n)$ to a weighted mixture $\{(x_n, \ell, Q(y_n, \ell))\}_{\ell=1}^{K}$ with $C$.

3. Apply regular classification algorithm on the weighted mixtures $\bigcup_{n=1}^{N} \{(x_n, \ell, Q(y_n, \ell))\}_{\ell=1}^{K}$.

- Mixture uncertainty: entropy of each normalized $Q(y, :)$
- A simple and unique optimal shifting exists for every $C$
  $$Q(y, k) = \max_{\ell} C(y, \ell) - C(y, k)$$

Good $g$ for new regular classification problem $\Rightarrow$ Good $g$ for cost-sensitive classification problem.
Unavoidable (Minimum) Uncertainty

Original Cost-Sensitive Classification Problem

individual examples with certainty

New Regular Classification Problem

mixtures with unavoidable uncertainty

new problem usually **harder** than original one

*need robust* regular classification algorithm to deal with uncertainty
One-Versus-One: A Popular Classification Meta-Method

1. For a pair \((i, j)\), take all examples \((x_n, y_n)\) that \(y_n = i\) or \(j\).
2. Train a binary classifier \(g^{(i,j)}\) using those examples.
3. Repeat the previous two steps for all different \((i, j)\).
4. Predict using the votes from \(g^{(i,j)}\).

Cost-sensitive multiclass classification → regular (weighted) multiclass classification

OVO decomposition → regular (weighted) binary classification

**cost-sensitive one-versus-one:**

cost transformation + one-versus-one
Cost-Sensitive One-Versus-One (CSOVO)

1. for a pair \((i, j)\), transform all examples \((x_n, y_n)\) to
   \[ (x_n, \arg\min_{k \in \{i, j\}} C(y_n, k)) \]
   with weight \(|C(y_n, i) - C(y_n, j)|\)

2. train a binary classifier \(g^{(i,j)}\) using those examples

3. repeat the previous two steps for all different \((i, j)\)

4. predict using the votes from \(g^{(i,j)}\)

comes with **good theoretical guarantee**:

\[
\text{test cost of final classifier} \leq 2 \sum_{i < j} \text{test cost of } g^{(i,j)}
\]

**simple, efficient**, and
takes original OVO as **special case**
OVO: popular regular classification meta-method, NOT cost-sensitive

couple both meta-methods with SVM

CSOVO often better suited for cost-sensitive classification
• **cost transformation** methodology: makes any (robust) regular classification algorithm cost-sensitive
• theoretical guarantee: **cost equivalence**
• algorithmic use: a **novel and simple** algorithm CSOVO
• experimental performance of CSOVO: **superior**

Thank you for your attention!