Estimating the Class Probability Threshold without Training Data

Ricardo Blanco-Vega,
César Ferri,
José Hernández-Orallo,
María José Ramírez-Quintana

Dep. de Sist. Informàtics i Computació, Universitat Politècnica de València, Spain

Third workshop on ROC Analysis, ROCML’06
Introduction

- ROC analysis: The traditional solution to the problem of contextualising a classifier to a new cost
- In order to perform ROC analysis (as well as other techniques), we need a training or validation dataset
  - In some situations, however, we don't have any training or validation data analysis available.
Introduction

- We face this situation:
  - When we have to adapt an existing method which was elaborated by a human expert
  - We do not have the old training data which was used to construct the initial model available

- Traditional techniques cannot be applied if we do not have the original data
Introduction

- The mimetic technique generate comprehensible models from non-comprehensible models by imitating its behaviour.

- In this work we use this technique in order to adapt old models to new cost contexts.

- Additionally, we are interested in the comprehensibility of the new model.
If we employ a very huge invented dataset, the new model has a fidelity close to 100% wrt the original model.
Mimetic Technique

- In this paper we research how to adapt the mimetic technique
Mimetic Technique

- In this paper we research how to adapt the mimetic technique:
  - Threshold estimation
  - Invented dataset generation
  - Mimetic schema
Threshold Estimation

- We also study three different methods of estimating the optimal threshold:
  - Direct threshold
  - Ranking threshold
  - ROC threshold
Threshold Estimation

- **Direct Method (Elkan, 2001):**

\[
\text{skew} = \frac{C(0,1) - C(1,1)}{C(1,0) - C(0,0)} \quad p^* = \frac{1}{1 + \text{skew}}
\]

\[
\text{Threshold}_{\text{Dir.}} = 1 - p^* = \frac{\text{skew}}{1 + \text{skew}}
\]
Threshold Estimation

- Ranking threshold:
  - The idea is to employ the estimated probabilities directly to compute the threshold.
  - After ranking the examples, we select a point between two points in this rank such that there are (approximately) \( \frac{n}{\text{skew}+1} \) examples on the left side and \( (n^* \text{skew}/(\text{skew}+1)) \) examples on the right side.
Threshold Estimation

- **ROC threshold:**
  - Suppose that a model is well calibrated
    - If a model gives a probability 0.8 of being class 0 to 100 examples, 80 should be of class 0, and 20 should be of class 1
    - In the ROC space, this will be a segment going from point (0,0) to the point (20,80) with a slope of 4
  - We define a version of the ROC curve named NROC based on the idea that a probability represents a percentage of correctly classified instances (calibrated classifier)
Threshold Estimation

- ROC threshold:

\[ N^0 = \{1, 0.9, 0.7, 0.4\} \quad \text{Sum}^0 = 3 \quad N^0 = \{0.33, 0.3, 0.23, 0.13\} \quad N^0 = \{0.33, 0.63, 0.86, 1\} \]

\[ N^1 = \{0, 0.1, 0.3, 0.6\} \quad \text{Sum}^1 = 1 \quad N^1 = \{0, 0.1, 0.3, 0.6\} \quad N^1 = \{0, 0.1, 0.4, 1\} \]

\[
\begin{align*}
(0, 0) \\
(0, 0.33) \\
(0.1, 0.63) \\
(0.4, 0.86) \\
(1, 1)
\end{align*}
\]
Threshold Estimation

- **ROC threshold:**
  - Since we work on a normalised ROC space we need to normalise the skew:
    \[
    skew' = skew \cdot \frac{\text{Sum}^0}{\text{Sum}^1}
    \]
  - If skew' is exactly parallel to a segment, then the threshold must be exactly the probability that corresponds to that segment:
    \[
    \text{Threshold}_{ROC} = \frac{skew'}{1 + skew'}
    \]
Threshold Estimation

- **ROC threshold:**
  - **Simplifying:**

\[
\text{Threshold}_\text{ROC} = \frac{\text{skew} \cdot \frac{\text{Sum}^0}{\text{Sum}^1}}{1 + \text{skew} \cdot \frac{\text{Sum}^0}{\text{Sum}^1}}
\]

\[
\text{Threshold}_\text{ROC} = \frac{1}{1 + \frac{1}{\text{skew}} \cdot \frac{\text{Sum}^0}{\text{Sum}^1}}
\]
Threshold Estimation

- **Threshold Comparison:**
  - **Maximum:**
    - For the direct and the ROC methods, the maximum (1) is obtained when skew=∞:
    - The upper limit of ThresholdOrd is given by the example with highest probability.
  - **Minimum:**
    - For the direct and the ROC methods, the minimum (0) is obtained when skew=0
    - The lower limit of ThresholdOrd is given by the example with lowest probability
Threshold Estimation

- **Threshold Comparison:**
  - We can found cases for which \( \text{ThresholdDir} > \text{ThresholdOrd} \), and vice versa.
  - The relationship between \( \text{ThresholdROC} \) and \( \text{ThresholdDir} \) depends on the relationship between \( \text{Sum1} \) and \( \text{Sum0} \).
  - The three thresholds **coincide** when the probabilities are uniformly distributed.
Invented Dataset

- We study three methods for generating the invented dataset $D$:
  - **A priori method**: $D$ preserves the class distribution of the original training dataset
  - **Balanced method**: The same number of examples of each class is generated by this method
  - **Random method**: no conditions about the class frequency in $D$ are imposed
  - **Oversampling method**: $D$ contains a proportion of $1/(skew+1)$ of instances of class 0 and a proportion of $skew/(skew+1)$ of instances of class 1.
Mimetic Learning Schemas

- **Scheme 0 (Mim0):** The context is not taken into account.

```plaintext
D → Ω → D_l → J48 with pruning → μ → Threshold = 0.5
```
Mimetic Learning Schemas

- **Scheme 1 (Mim1):** a posteriori scheme, the context information is used when the mimetic model is applied.

```plaintext
D → Ω → D' → J48 without pruning → μ → Threshold T Calculation → Threshold = T
```

Cost Matrix
Mimetic Learning Schemas

- **Scheme 2 (Mim2):** a priori scheme in which the context information is used before the mimetic model is learned (MetaCost)
**Mimetic Learning Schemas**

- **Scheme 3 (Mim3):** the context information is used for generating the invented dataset using oversampling.
Experiments

- 22 Configurations
  - **Mim1 + Mim2** (18)
    - 3 Threshold methods
    - 3 Methods for generating $D$
  - **Mim0** (3)
    - 3 Methods for generating $D$
  - **Mim3** (1)
Experiments

- 10x10-fold cross-validation
- 20 binary datasets from the UCI repository
  - 10 Balanced
  - 10 Unbalanced
- J48 + Laplace Correction for learning the mimetic model
- NaiveBayes and ANN for the oracle
- The size of the invented dataset is 10,000
## Experiments

<table>
<thead>
<tr>
<th>Model</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(5)</th>
<th>(10)</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mim0a</td>
<td>19.74</td>
<td>29.58</td>
<td>39.68</td>
<td>59.53</td>
<td>109.81</td>
<td>51.67</td>
</tr>
<tr>
<td>Mim0b</td>
<td>20.24</td>
<td>30.30</td>
<td>40.27</td>
<td>60.91</td>
<td>114.45</td>
<td>53.24</td>
</tr>
<tr>
<td>Mim0c</td>
<td>20.56</td>
<td>31.35</td>
<td>41.95</td>
<td>63.39</td>
<td>116.87</td>
<td>54.82</td>
</tr>
<tr>
<td>Mim1Dira</td>
<td>20.00</td>
<td>29.41</td>
<td>38.03</td>
<td>51.34</td>
<td>75.88</td>
<td>42.93</td>
</tr>
<tr>
<td>Mim1Dirb</td>
<td>20.37</td>
<td>30.05</td>
<td>38.37</td>
<td>51.19</td>
<td>73.11</td>
<td>42.62</td>
</tr>
<tr>
<td>Mim1Dirc</td>
<td>20.72</td>
<td>30.82</td>
<td>39.29</td>
<td>51.74</td>
<td>73.16</td>
<td>43.15</td>
</tr>
<tr>
<td>Mim1Orda</td>
<td>25.20</td>
<td>32.00</td>
<td>34.32</td>
<td>36.86</td>
<td>39.07</td>
<td>33.49</td>
</tr>
<tr>
<td>Mim1Ordb</td>
<td>20.67</td>
<td>29.36</td>
<td>33.03</td>
<td>36.43</td>
<td>38.76</td>
<td>31.65</td>
</tr>
<tr>
<td>Mim1Ordc</td>
<td>26.17</td>
<td>34.85</td>
<td>39.49</td>
<td>43.84</td>
<td>47.22</td>
<td>38.32</td>
</tr>
<tr>
<td>Mim1ROCa</td>
<td>20.25</td>
<td>29.65</td>
<td>37.80</td>
<td>51.03</td>
<td>71.92</td>
<td>42.13</td>
</tr>
<tr>
<td>Mim1ROCb</td>
<td>20.37</td>
<td>29.84</td>
<td>37.66</td>
<td>50.60</td>
<td>72.49</td>
<td>42.19</td>
</tr>
<tr>
<td>Mim1ROCc</td>
<td>20.73</td>
<td>29.31</td>
<td>36.54</td>
<td>48.25</td>
<td>68.92</td>
<td>40.75</td>
</tr>
<tr>
<td>Mim2Dira</td>
<td>19.74</td>
<td>27.00</td>
<td>31.86</td>
<td>39.25</td>
<td>51.74</td>
<td>33.92</td>
</tr>
<tr>
<td>Mim2Dirb</td>
<td>20.24</td>
<td>27.58</td>
<td>32.24</td>
<td>39.67</td>
<td>51.91</td>
<td>34.33</td>
</tr>
<tr>
<td>Mim2Dirc</td>
<td>20.56</td>
<td>29.02</td>
<td>34.69</td>
<td>42.98</td>
<td>58.87</td>
<td>37.22</td>
</tr>
<tr>
<td>Mim2Orda</td>
<td>23.80</td>
<td>29.98</td>
<td>32.75</td>
<td>35.77</td>
<td>37.42</td>
<td>31.94</td>
</tr>
<tr>
<td>Mim2Ordb</td>
<td>20.61</td>
<td>28.51</td>
<td>31.69</td>
<td>34.84</td>
<td>37.03</td>
<td>30.54</td>
</tr>
<tr>
<td>Mim2Ordc</td>
<td>24.75</td>
<td>33.26</td>
<td>38.11</td>
<td>41.83</td>
<td>45.74</td>
<td>36.74</td>
</tr>
<tr>
<td>Mim2ROCa</td>
<td>20.43</td>
<td>27.77</td>
<td>32.65</td>
<td>39.10</td>
<td>50.22</td>
<td>34.03</td>
</tr>
<tr>
<td>Mim2ROCb</td>
<td>20.58</td>
<td>28.13</td>
<td>33.75</td>
<td>40.83</td>
<td>53.47</td>
<td>35.35</td>
</tr>
<tr>
<td>Mim2ROCc</td>
<td>20.83</td>
<td>29.30</td>
<td>34.92</td>
<td>44.46</td>
<td>60.12</td>
<td>37.93</td>
</tr>
<tr>
<td>Mim3</td>
<td>20.33</td>
<td>28.51</td>
<td>34.85</td>
<td>44.69</td>
<td>61.65</td>
<td>38.01</td>
</tr>
</tbody>
</table>
### Experiments

<table>
<thead>
<tr>
<th>Model</th>
<th>Balanced</th>
<th>Majority</th>
<th>Minority</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mim0a</td>
<td>77.40</td>
<td>23.32</td>
<td>73.62</td>
</tr>
<tr>
<td>Mim0b</td>
<td>77.53</td>
<td>23.46</td>
<td>72.09</td>
</tr>
<tr>
<td>Mim0c</td>
<td>76.24</td>
<td>29.53</td>
<td>72.92</td>
</tr>
<tr>
<td>Mim1DirA</td>
<td>76.22</td>
<td>20.30</td>
<td>73.36</td>
</tr>
<tr>
<td>Mim1DirB</td>
<td>76.32</td>
<td>20.34</td>
<td>73.06</td>
</tr>
<tr>
<td>Mim1DirC</td>
<td>75.62</td>
<td>22.98</td>
<td>72.98</td>
</tr>
<tr>
<td>Mim1OrdA</td>
<td>65.38</td>
<td>17.39</td>
<td>69.53</td>
</tr>
<tr>
<td>Mim1OrdB</td>
<td>65.49</td>
<td>17.44</td>
<td>70.40</td>
</tr>
<tr>
<td>Mim1OrdC</td>
<td>63.88</td>
<td>20.87</td>
<td>68.03</td>
</tr>
<tr>
<td>Mim1RocA</td>
<td>76.31</td>
<td>20.28</td>
<td>73.42</td>
</tr>
<tr>
<td>Mim1RocB</td>
<td>76.37</td>
<td>20.29</td>
<td>73.29</td>
</tr>
<tr>
<td>Mim1RocC</td>
<td>75.81</td>
<td>19.52</td>
<td>73.36</td>
</tr>
<tr>
<td>Mim2DirA</td>
<td>75.05</td>
<td>14.59</td>
<td>74.11</td>
</tr>
<tr>
<td>Mim2DirB</td>
<td>75.06</td>
<td>14.59</td>
<td>73.37</td>
</tr>
<tr>
<td>Mim2Dire</td>
<td>73.75</td>
<td>20.96</td>
<td>73.33</td>
</tr>
<tr>
<td>Mim2OrdA</td>
<td>67.58</td>
<td>16.53</td>
<td>71.32</td>
</tr>
<tr>
<td>Mim2OrdB</td>
<td>67.64</td>
<td>16.50</td>
<td>71.56</td>
</tr>
<tr>
<td>Mim2OrdC</td>
<td>65.98</td>
<td>20.39</td>
<td>69.82</td>
</tr>
<tr>
<td>Mim2ROCa</td>
<td>75.64</td>
<td>15.50</td>
<td>73.86</td>
</tr>
<tr>
<td>Mim2ROCb</td>
<td>75.57</td>
<td>15.65</td>
<td>73.21</td>
</tr>
<tr>
<td>Mim2ROCc</td>
<td>74.73</td>
<td>20.03</td>
<td>72.81</td>
</tr>
<tr>
<td>Mim3</td>
<td>76.08</td>
<td>19.44</td>
<td>73.38</td>
</tr>
<tr>
<td>Oracle</td>
<td>81.43</td>
<td>20.61</td>
<td>77.57</td>
</tr>
</tbody>
</table>
Experiments

- Mim2 schema presents the best behaviour

- For the generation of the invented dataset, $a$ (a priori) and $b$ (balanced) are clearly better than $c$ (random)

- ThresholdOrd option seems to give the best results w.r.t. costs
Conclusions

- Not having data is not an obstacle if we want to adapt an existing model

- The introduced techniques are useful to reduce the costs of the model, overcoming classical approaches (oversampling)

- Mim2 (a priori) obtains the best performance
Conclusions

- We have presented several methods to derive a class threshold without training or validation data

  - The approach based on sorting the probabilities only assumes that the probabilities are reasonably well ordered

  - The approach based on ROC analysis is optimal if the probabilities are well calibrated
Future Work

- Analyse the threshold derivation methods after performing a calibration

- Hybrid techniques between the Ord and ROC methods