Reliable Calibrated Probability Estimation in Classification

MLDM Workshop

Marinka Zitnik
http://helikoid.si
Mentor: Assoc. Prof. Marko Robnik Sikonja, Ph. D.
Final Project Presentation

January 17, 2012
Motivation

▶ Probability calibration: Expected probability of correct guesses differs from real proportions.

▶ Reliable probability and confidence estimation is crucial in many applications: (i) assessment of risks and costs, (ii) robust integration with background knowledge, (iii) multi classifiers – bad local estimations distort the global result.

▶ Binning method and isotonic regression are often used calibration methods.

▶ **Problem:** Binning and isotonic regression give poor results on small and noisy calibration sets.
  ▶ Inappropriate intervals (too many or too few bins).
  ▶ Boundary generalization (poor performance on the edges of bins).
  ▶ Errors in examples distribute to all examples in the section.

▶ **Solution:** Calibrate with isotonic regression and binning method by using bootstrapping technique (**Boot-Binning, Boot-Isotonic Regression**). Use confidence intervals for merging unreliable or too narrow calibration intervals.
Univariate Calibration Methods

- **Simple Normalization**
  - (Non) linear transformations for pre-calibration.
  - E-calibration, Softmax calibration.

- **Calibration Using Mapping**
  - Mapping function from membership value $p_i$ to a calibrated conditional probability for positive class $\hat{f}_i$.
  - Binning, Platt’s logistic regression, Piecewise (Full) logistic regression, Isotonic regression.

- **Calibration via Bayes Rule**
  - Positive unnormalized class scores are split into two groups according to their true class. Membership probabilities are determined with Bayes theorem to class conditional probabilities and class priors.
  - Choice of the distribution type for the class conditional probabilities: Gaussian, Laplacian.

- **Calibration Using Assignment Values**
  - Membership values are partitioned according to their assignment. Modelled separately in each partition as Beta distributed random variables.
Calibration Quality Measures

- (i) mean squared error (MSE), (ii) LogLoss, (iii) calibration by overlapping bins (CalBin), (iv) calibration loss (CalLoss), (v) H-hat and (vi) chi-squared test through Hosmer-Lemeshow C-hat.

- **Pure** measures: CalBin, CalLoss.

- Brier score was originally not a calibration measure. It was decomposed in terms of calibration and refinement loss.

- **Impure** measures: MSE, LogLoss.

- **Insensitive to calibration**: qualitative – CA, F-measure; ranking – AUC.
Boot-Binning and Boot-Isotonic Regression

1. Construct bootstrapped data set:
   ▶ training set (sampled examples, 63%),
   ▶ calibration set (37%).

2.1 Classification model.

2.2 Calibration. Binning (Boot-Binning)/isotonic regression (Boot-Isotonic Regression).

3 Repeat 1 and 2 \( r \) times. Record calibrated probabilities.

4 Final calibration.
   ▶ Non parametric confidence interval and a pseudomedian estimate for each example \( x_i \) with Wilcoxon signed-rank test.
   ▶ Continually merge two sets of intervals obtained from 3. Probabilities are merged accordingly.

   ▶ Parametric methods. (i) \( r \), the number of iterations; (ii) \( T_{count} \), confidence interval threshold; (iii) \( T_{in} \), intersection between confidence interval and an individual bin.
Experimental Configuration


- **5 classifiers:** (i) decision trees (DT), (ii) support vector machines (SVM), (iii) random forests (RF), (iv) boosting (Boost), and (v) naive Bayes (NB).

- **5 calibration methods:** (i) binning (Binn), (ii) isotonic regression (IR), (iii) boot-isotonic regression (B-IR) and (iv) boot-binning (B-Binn), (v) Base.

- **2 quality measures:** CalBin and MSE.

- 30 repetitions for each combination of data set, classification, calibration – (7500 = 30 · 10 · 5 · 5) tests. Final CalBin and MSE measures averaged over 30 repetitions.
Experimental Evaluation

- Differences exist in results, evaluated with CalBin/MSE measure. This arises from the different nature of the measures.

- Observations confirm previous knowledge.
  - Isotonic regression is more powerful than binning. Binning and boot-binning methods are outperformed by isotonic regression and boot-isotonic regression.
  - Substantial differences between learning algorithms.

- Two-sided paired Wilcoxon signed-rank test for comparison of two classifiers over multiple data sets.

- Null hypothesis: Binning (isotonic regression) and boot-binning (boot-isotonic regression) perform equally well.

- 10 hypotheses: 5 classifiers and 2 proposed methods.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Binning</th>
<th>Isotonic regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT</td>
<td>0.00195; [0.0043, 0.0205]</td>
<td>0.03711; [0.0015, 0.0288]</td>
</tr>
<tr>
<td>SVM</td>
<td>0.00977; [0.0026, 0.0158]</td>
<td>0.06251; [0.0195, 0.0488]</td>
</tr>
<tr>
<td>RF</td>
<td>0.01898; [-0.0103, -0.0008]</td>
<td>0.05413; [0.0076, 0.0254]</td>
</tr>
<tr>
<td>Boost</td>
<td>0.02734; [0.0021, 0.0349]</td>
<td>0.04409; [0.0010, 0.0329]</td>
</tr>
<tr>
<td>NB</td>
<td>0.01953; [0.0010, 0.0158]</td>
<td>0.03091; [0.0019, 0.0291]</td>
</tr>
</tbody>
</table>

Table: Boot-Binning. $\alpha = 0.05$. Table: Boot-Isotonic reg. $\alpha = 0.05$. 
Anomalies Detection and Removal

- Evaluation on 6 artificial data sets. A. Gradual addition of noise. B. Proportion of positive class increases linearly with respect to the probability estimate + added class labels for 1.

**Figure:** Left. Reliability diagram for (boot-) isotonic regression for A. Right. Reliability diagram for (boot-) binning for B.
The problems of poor calibration results on small or noisy calibration sets have been eliminated.

Calibration results show significant improvement using the bootstrapping approach.

Null hypotheses, that binning (isotonic regression) and boot-binning (boot-isotonic regression) perform equally well is rejected for majority of classifiers.

Unable to reject the null hypotheses of equal performance between isotonic regression and its bootstrapping improvement for SVMs and random forests.

Future work.

- Multi class problems.
- Reduce complexity due to careful choice of parameters.