Smooth Isotonic Regression: A New Method to Calibrate Predictive Models

Xiaoqian Jiang, PhD[1], Melanie Osl, PhD[1], Jihoon Kim, MSc[1], Lucila Ohno-Machado, MD, PhD[1]
[1] Division of Biomedical Informatics, Department of Medicine
University of California, San Diego

Abstract

Predictive models are critical for risk adjustment in clinical research. Evaluation of supervised learning models often focuses on predictive model discrimination, sometimes neglecting the assessment of their calibration. Recent research in machine learning has shown the benefits of calibrating predictive models, which becomes especially important when probability estimates are used for clinical decision making. By extending the isotonic regression method for recalibration to obtain a smoother fit in reliability diagrams, we introduce a novel method that combines parametric and non-parametric approaches. The method calibrates probabilistic outputs smoothly and shows better generalization ability than its ancestors in simulated as well as real world biomedical data sets.

Introduction

Risk assessment tools such as the Cox proportional hazard model, the logistic regression model, and other machine-learning based predictive models are widely used in patient diagnosis, prognosis and clinical studies. Accurate calibration of these models is important if the outputs are going to be applied to new cohorts [3]. For example, the Gail model, a predictive model of a woman's risk of developing breast cancer, was reported to underestimate the risk among a specific subgroup of patients [8]. After recalibration, the model identified more patients who would benefit from chemoprevention than the original model [1]. Another example is derived from the Framingham Heart Study model, in which gender-specific coronary heart disease (CHD) prediction functions can be used for assessing the risk of developing CHD. While the original model overestimated the risk of 5-year CHD events among Japanese American men, Hispanic men and Native American women, the recalibrated risk score based on the new cohort's own average incidence rate, performed well [4].

A well calibrated predictive model provides risk estimates that reflect the underlying probabilities for an disease. This means that the proportion of positive events ($c = 1$ from $c \in \{0, 1\}$) in a group of cases that have according to the model a risk of e.g. $p = 0.8$ is exactly 0.8. Needless to say, this notion of calibration depends on an sufficient number of cases with the same risk to be evaluated reliably. In practice, when there are not many cases with the same estimated probability, cases with similar values for p are grouped for evaluation.

Calibration Assessment A simple way of assessing the calibration of a predictive model is a calibration plot or reliability diagram. This visual tool plots expected versus observed events as follows: all estimated probabilities p are grouped according to the fixed cutoff points 0.1, 0.2, ..., 1.0. The x-coordinates of the points in the plot are the mean values of the estimated probabilities in each group. The y-coordinates are the observed fraction of cases with $c = 1$. If the predictive model is well calibrated, the points fall near the diagonal line. An example of a calibration plot is given in Figure 1. The points of the plot are connected by a line for better visualization. The meaning of the dotted red lines is explained in the next paragraph.

A quantitative measure of calibration is given by a goodness-of-fit test like the well-known Hosmer-Lemeshow test [9]. The Homer-Lemeshow C-statistic is given by:

$$HL_c = \sum_{c=0}^{1} \sum_{i=1}^{G} \frac{(O_i^c - E_i)^2}{E_i(1 - \frac{E_1}{n_i})},$$

where G denotes the number of groups (usually 10), O_i^c is the sum of cases with $c = 0$ or $c = 1$, E_i is the sum of estimated probabilities, and n_i denotes the number of cases in group i. This value is then compared to a chi-square distribution with G-2 degrees of freedom.

To improve the calibration of binary classification models, different calibration methods have been proposed.

Calibration Improvement Two popular methods to improve the calibration of predictive models are the

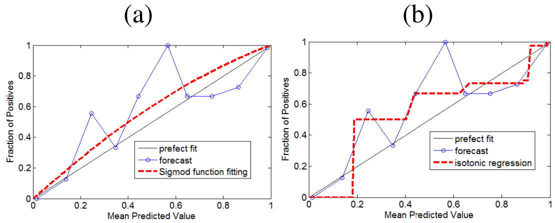

Figure 1: Calibration plot with fitted probabilities by (a) sigmoid fitting and (b) isotonic regression.

methods proposed by Platt [13] and Zadrozny and Elkan [15].

The parametric approach of Platt applies to model output rank p consists of finding the parameters A and B for a sigmoid function $f_s(p) = \frac{1}{1+e^{A \cdot p+B}}$, such that the negative log-likelihood is minimized. However, this method may lead to poor calibration when the outputs do not fit the sigmoid function. The application of this method, referred to as sigmoid fitting, is shown in Figure 1 (a) by the dotted red line.

The non-parametric approach of Zadrozny and Elkan applies a pair-adjacent violators algorithm [2] to the previously sorted output probabilities p of the model in order to find a stepwise-constant isotonic function that best fits according to a mean-squared error criterion. However, the outputs of this calibration method tend to overfit the data if no smoothing regularization is applied. The use of this method, referred to as isotonic regression, is shown in Figure 1 (b) by the dotted red line.

Smooth non-parametric estimators are expected to alleviate overfitting and underfitting problems, and thus have received more attention recently. The methods by Wang et al. [14] and Meyer [10] find a non-decreasing mapping function $t()$ that minimizes:

$$\sum_i (c_i - t(p_i))^2 + \lambda \int_a^b [t^{(m)}(\gamma)]^2 \, d\gamma, \qquad (1)$$

where m corresponds to a smoothness parameter, a and b represent the range of input predictions, and λ balances the goodness-of-fit (first component) and the smoothness (second component) of the transformation function $t()$. When $m = 1$, Equation 1 corresponds to a piece-wise linear estimator. When $m = 2$, Equation 1 represents a smooth monotone estimator. In theory, these models are smoother than isotonic regression and more flexible than sigmoid regression, but their inferences require much heavier computation and tedious parameter tuning. Moreover, approximation algorithms show large empirical losses, e.g. 30% using second-order cone programming [14].

Method

We intended to develop a smoother yet computationally affordable method to further improve the calibration of predictive models. We observed that isotonic regression is a non-parametric method that joins predictions into larger bins, as indicated by the flat regions in Figure 1(b). By interpolating between a few representative values, we can obtain a smoother function. However, we must ensure that such interpolation function $g()$ is monotonically increasing to maintain the discriminative ability of the predictive model. Let $P = \{p_i\}$ the set of all predictions p_i and $C = \{c_i\}$ their corresponding class labels, then the function $t^*() = g(f(P'), C')$ is also monotonically increasing, where $f()$ is the isotonic regression function, P' and C' are subsets of predictions and their corresponding class labels, repectively.

Based on these considerations, we propose a novel approximation to the optimal smooth function $t^*()$ that minimizes Equation 1 in three steps. First, we apply isotonic regression to obtain a monotone non-parametric function $f()$ that minimizes $\sum_i (c_i - f(p_i))^2$. Second, we select s representative points from the isotonic mapping function. Finally, we construct a monotonic spline, called Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) [7], that interpolates between the sampled points from the isotonic regression function to obtain a smoothed approximation to Equation 1. Note that the interpolation by PCHIP is monotonic to keep the partial ordering of the predictive model probabilities, and that it introduces the smoothness. Algorithm 1 gives the details on these steps in a compact form.

Inputs: Prediction probabilities $P = p_1, \ldots, p_n$, class labels $C = c_1, \ldots, c_n$.
Output: Smoothed isotonic regression function h.

1. Obtain $f^* = \text{argmin}_f \sum_i (c_i - f(p_i))^2$, subject to $f(p_i) \leq f(p_{i+1}), \forall i$ (Isotonic Regression).
2. Sample s points from $f^*(\gamma), \gamma \in (0, 1)$, one point per flat region of $f^*(\cdot)$. Denote samples as P', and their corresponding class labels as C'.
3. Construct a Piecewise Cubic Hermite Interpolating Polynomial function $t^*(f^*(P'), C'))$ to obtain a monotone smoothing spline as the final transformation function for calibration.

Algorithm 1: Smooth Isotonic Regression

Experiments

We compare our method with sigmoid fitting and isotonic regression on the task of improving the calibration of logistic regression (LR) models learned on synthetic and real world data.

Synthetic Data We took random samples of size $n = 1000$ from two Gaussian distributions with varying differences in means but fixed variances. The differences between μ_1 and μ_2 were set to $0.5, 1.0, 1.5$ and 2.0 and the variances were set to $\Sigma_1 = 2.0, \Sigma_2 = 1.0$, respectively. We used 80% of the generated data sets to train the LR model, and 20% to test the calibration of the predictions from the LR model and the predictions after recalibration by sigmoid fitting, logistic regression, and our method.

The results of this experiment are shown in Figure 2. The blue circles in the plots are the predicted probabilities of the LR model. The red dotted lines are the recalibrated probabilities. While sigmoid fitting does not improve the calibration in all four cases, both isotonic regression and smooth isotonic regression follow the data pattern closely. They smooth isotonic regression has less oscillation and has a p-value larger than 0.05 for the H-L test indicating that the recalibrated predictions are reasonably well calibrated. We further observe that isotonic regression tends to overfit, while smooth isotonic regression provides a continuous recalibration and the highest p-values for the HL-test in most cases.

Real World Experiment We used eight different real world data sets. GSE2034 and GSE2990 are gene expression data sets related to breast cancer. Both data sets were preprocessed to keep only the top 15 features (see [12] for details). The HOSPITAL data consists of microbiology cultures and other variables related to hospital discharge errors of a subgroup in [5]. ADULT, BANKRUPTCY, HEIGHT_WEIGHT, MNISTALL, PIMATR were obtained from the UCI Repository [6]. MNISTALL has handwritten numbers '0-9'. The problem has been converted into a binary problem by treating all digits '0' as positive and the others as negative, yielding a very unbalanced set. For each data set, we learned an LR model on 60% random samples and tested on the remaining 40%, with the exception of the ADULT dataset, where we followed the split used in [11]. A summary of the data sets is given in Table 1. The percentage of positive cases varies from 8% to 67%.

Data	# Attr	Train size	Test size	% POS
GSE2034	15	125	84	54
GSE2990	15	54	36	67
ADULT	14	4,000	41,222	25
BANKRUPTCY	2	40	26	48
HEIGHT_WEIGHT	2	126	84	64
HOSPITAL	22	2,891	1,927	8
MNISTALL	784	42,000	28,000	9.8
PIMATR	8	120	80	33

Table 1: Real world data sets used. % POS indicates the percentage of positive cases.

Figure 3 shows histograms of the predicted values (top row) and calibration plots for the predictions of logistic regression, after sigmoid fitting, isotonic regression, and smooth isotonic regression on all eight test sets. None of the calibration methods decreases the AUC, since the monotonic transformation functions preserve the orderings. Isotonic regression sometimes shows an increase in AUC because it introduces more ties into the ranking.

An interesting observation gathered from the calibration plots is that they seldom display a sigmoid shape. Because the result is mostly data-driven, it discourages the use of a sigmoid function to transform predictions into probabilities (see third row). The calibration plots in the fourth row of the figure show results for isotonic regression, which are not smooth and are unrealistically sharp at the corners. The calibration plots at the bottom of the figure show the functions fitted with our proposed smooth isotonic regression, which have better performance than sigmoid fitting and less oscillation than isotonic regression. In all cases, smooth isotonic regression gives the highest p-value for the HL-test suggesting a better fit than the sigmoid approach and less overfit when compared to isotonic regression.

Conclusion

There is increasing interest in improving the calibration of predictive models, especially given their potential use for personalized medicine. While discrimination is often optimized, calibration is sometimes neglected, potentially leading to the publication of models that are not adequate for use in practice. We proposed a smooth isotonic regression method that significantly improves simple isotonic regression. The method combines the merits of parametric and non-parametric models, providing a smooth non-parametric method to improve the calibration of predictive models.

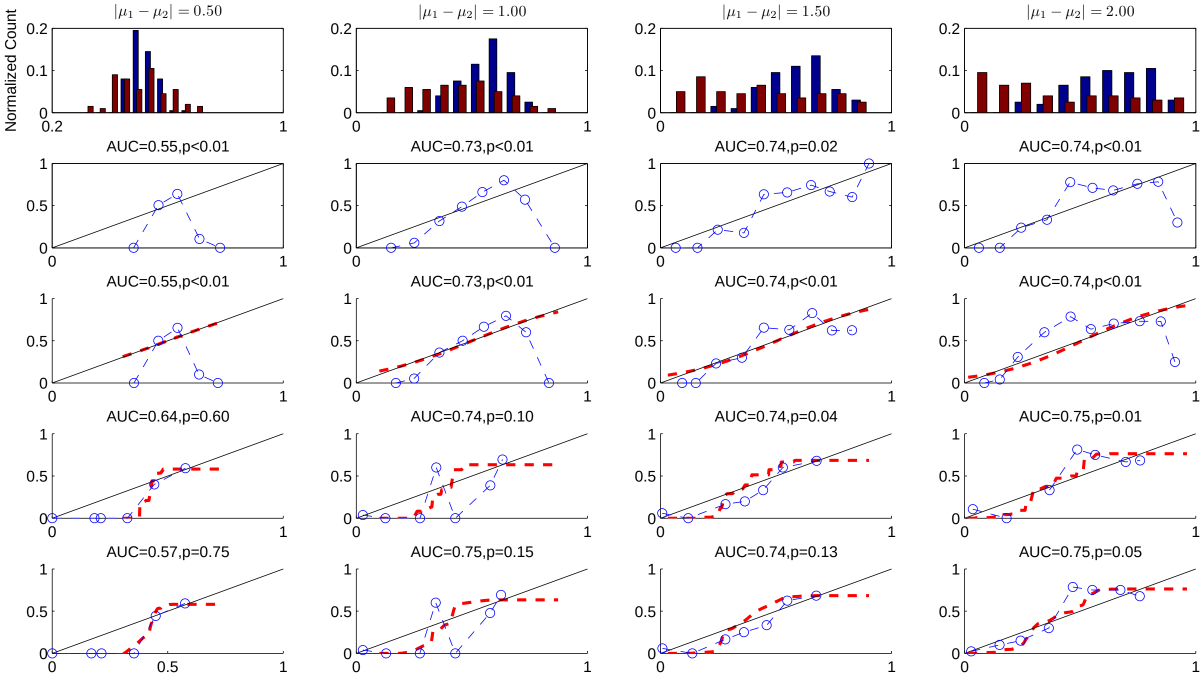

Figure 2: Comparison of different calibration methods on synthetic data. Row one shows histograms of the original predicted probabilities by LR (blue bars for class $c = 0$ and red bars for class $c = 1$). Row two to five show calibration plots for the originall predicted probabilities of LR and the recalibrated probabilities after sigmoid fitting, isotonic regression, and smooth isotonic regression. The caption of each figure contains the discriminatory ability in terms of the area under the ROC curve (AUC) and the p-value of the HL test for the visualized probabilities.

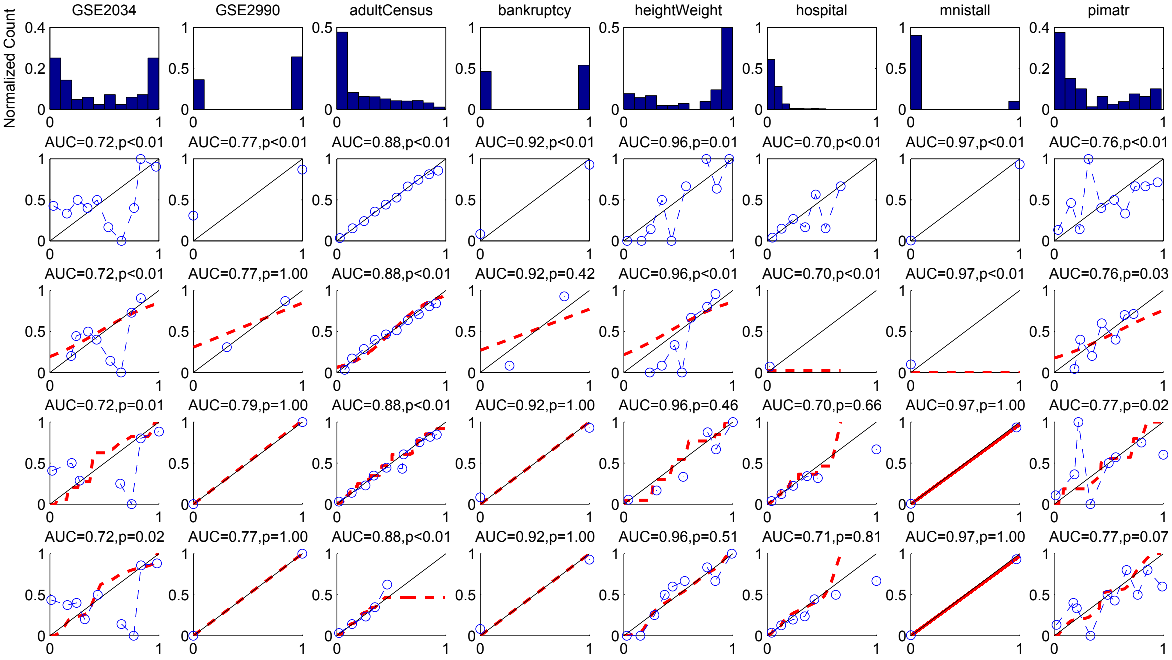

Figure 3: Comparison of different calibration methods on real world data. Row one shows histograms of the original predicted values by LR (no color discrimination for classes is used). Row two to five show calibration plots for the originally predicted probabilities of LR and the recalibrated probabilities after sigmoid fitting, isotonic regression, and smooth isotonic regression. The caption of each figure contains the discriminatory ability in terms of the area under the ROC curve (AUC) and the p-value of the HL test for the visualized probabilities.

Acknowledgements

This work was funded in part by the Komen Foundation (FAS0703850), the National Library of Medicine (R01LM009520), the Austrian Genome Program (GEN-AU), project Bioinformatics Integration Network (BIN), NHLBI (U54 HL10846), and R01 HS19913-01. We thank Dr. El-Kareh for making one of the data sets available for this study.

References

[1] E. Amir, O. C. Freedman, B. Seruga, and D. G. Evans. Assessing women at high risk of breast cancer: a review of risk assessment models. *J Natl Cancer Inst*, 102(10):680–691, 2010.

[2] M. Ayer, H. Brunk, G. Ewing, W. Reid, and E. Silverman. An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics*, 26(4):641–647, 1955.

[3] I. Cohen and M. Goldszmidt. Properties and benefits of calibrated classifiers. In *8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 125–136, 2004.

[4] R. B. D'Agostino, S. Grundy, L. M. Sullivan, and P. Wilson. Validation of the framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA*, 286(2):180–187, 2001.

[5] R. El-Kareh, C. Roy, G. Brodsky, M. Perencevich, and E. G. Poon. Incidence and predictors of microbiology results returning post-discharge and requiring follow-up. *Journal of Hospital Medicine*, 2010, (accepted).

[6] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[7] F. N. Fritsch and R. E. Carlson. Monotone piecewise cubic interpolation. *SIAM J Numer Anal*, 17(2):238–246, 1980.

[8] M. H. Gail, L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, C. Schairer, and J. J. Mulvihill. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natl. Cancer Inst.*, 81:1879–1886, Dec 1989.

[9] D. W. Hosmer, T. Hosmer, S. Le Cessie, and S. Lemeshow. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med*, 16(9):965–980, 1997.

[10] M. C. Meyer. Inference using shape-restricted regression splines. *Annals of Applied Statistics*, 2(3):1013–1033, 2008.

[11] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *International Conference on Machine Learning*, pages 625–632, 2005.

[12] M. Osl, S. Dreiseitl, J. Kim, K. Patel, and L. Ohno-Machado. Effect of data combination on predictive modeling: a study using gene expression data. In *AMIA Annual Symposium*, 2010.

[13] J. C. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. *Advances in Large Margin Classifiers*, pages 61–74, 1999.

[14] X. Wang and F. Li. Isotonic smoothing spline regression. *J Comput Graph Stat*, 17(1):21–37, 2008.

[15] B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699, 2002.