OPEN ACCESS Freely available online

PLOS ONE

Doubly Optimized Calibrated Support Vector Machine (DOC-SVM): An Algorithm for Joint Optimization of Discrimination and Calibration

Xiaoqian Jiang[1]*, Aditya Menon[2], Shuang Wang[1], Jihoon Kim[1], Lucila Ohno-Machado[1]

1 Division of Biomedical Informatics, University California San Diego (UCSD), La Jolla, California, United States of America, **2** Deptartment of Computer Science and Engineering, University California San Diego (UCSD), La Jolla, California, United States of America,

Abstract

Historically, probabilistic models for decision support have focused on discrimination, e.g., minimizing the ranking error of predicted outcomes. Unfortunately, these models ignore another important aspect, calibration, which indicates the magnitude of correctness of model predictions. Using discrimination and calibration simultaneously can be helpful for many clinical decisions. We investigated tradeoffs between these goals, and developed a unified maximum-margin method to handle them jointly. Our approach called, Doubly Optimized Calibrated Support Vector Machine (DOC-SVM), concurrently optimizes two loss functions: the ridge regression loss and the hinge loss. Experiments using three breast cancer gene-expression datasets (i.e., GSE2034, GSE2990, and Chanrion's datasets) showed that our model generated more calibrated outputs when compared to other state-of-the-art models like Support Vector Machine ($p = 0.03$, $p = 0.13$, and $p < 0.001$) and Logistic Regression ($p = 0.006$, $p = 0.008$, and $p < 0.001$). DOC-SVM also demonstrated better discrimination (i.e., higher AUCs) when compared to Support Vector Machine ($p = 0.38$, $p = 0.29$, and $p = 0.047$) and Logistic Regression ($p = 0.38$, $p = 0.04$, and $p < 0.0001$). DOC-SVM produced a model that was better calibrated without sacrificing discrimination, and hence may be helpful in clinical decision making.

Citation: Jiang X, Menon A, Wang S, Kim J, Ohno-Machado L (2012) Doubly Optimized Calibrated Support Vector Machine (DOC-SVM): An Algorithm for Joint Optimization of Discrimination and Calibration. PLoS ONE 7(11): e48823. doi:10.1371/journal.pone.0048823

Editor: Shree Ram Singh, National Cancer Institute, United States of America

Received May 15, 2012; **Accepted** October 3, 2012; **Published** November 6, 2012

Funding: XJ, JK and LOM were funded in part by the National Library of Medicine (1K99LM01139201, R01LM009520) and NHLBI (U54 HL108460). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: x1jiang@ucsd.edu

Introduction

Supervised learning has been widely applied in bioinformatics [1]. Given sufficient observations and their class memberships, the prediction task is often modeled by supervised learning algorithms, which aim at finding an optimal mapping between features and outcomes (usually represented by the zero-one class membership). In clinical predictions, *discrimination* measures the ability of a model to separate patients with different outcomes (e.g., positive or negative). In the case of a binary outcome, good *discrimination* indicates an adequate distinction in the distributions of predicted scores. That is, *discrimination* is determined by the degree of correct ranking performance of predicted scores [2]. On the other hand, *calibration* reflects the level to which observed probabilities match the predicted scores [3], e.g., the prediction average is 60% for every individual in a group of observations and the proportion of the positive observations is also 60% in that group.

Traditionally, many machine learning models were developed to optimize discrimination ability [4], (i.e., minimizing the errors in making binary decisions based on the model's estimates). However, in many direct-to-consumer applications (i.e., using molecular biomarkers for diagnostic or prognostic purposes [5,6]), estimated probabilities are being communicated directly to patients, hence calibration is very important. For example, clinicians may use estimated probabilities to make decisions related to prophylaxis for breast cancer. Achieving high levels of calibration in predictive models has become very important in clinical decision support and personalized medicine [7,8,9,10,11]."

Good *discrimination* may lead to good *calibration*, but this is not guaranteed. A highly *discriminative* classifier, i.e., one with a large Area Under the ROC Curve (AUC), might not necessarily be a calibrated one. Figure 1 illustrates an example with 20 simulated subjects. While two probabilistic model A and B have the same AUCs, the values of probabilities from model B are ten times smaller than those from model A. Although *discrimination* estimates the ranking of subjects and their class membership, it does not account for the consistency between probabilistic model predictions and the true underlying probabilities. In extreme cases, a classifier can draw a perfect decision boundary but produces unrealistic risk estimates (e.g., by estimating a probability of "0.01" for negative observations and "0.011" for positive observations). Thus, significant problems may occur when direct outputs of supervised classification models are blindly used as proxies to evaluate the "true risks".

In summary, although it is relatively easy to evaluate rank of estimates, it is non-trivial to convert these rankings into reliable probabilities of class membership, which is an important problem in personalized clinical decision making [12]. We want to find an accurate estimation of $p(y|X)$: the probability that a subject X belongs to class y, without sacrificing the *discriminative* ability of the

(a) Probabilistic classifier A

#	1	2	3	4	5	6	7	8	9	10
y	1	1	0	1	1	1	0	0	1	0
$\hat{y}$	.9	.8	.7	.6	.55	.54	.53	.52	.51	.505
#	11	12	13	14	15	16	17	18	19	20
y	1	0	1	1	0	0	1	0	1	0
$\hat{y}$	.4	.39	.38	.37	.36	.35	.34	.33	.3	.1

(b) Probabilistic classifier B

1	2	3	4	5	6	7	8	9	10
1	1	0	1	1	1	0	0	1	0
.09	.08	.07	.06	.055	.054	.053	.052	.051	.0505
11	12	13	14	15	16	17	18	19	20
1	0	1	1	0	0	1	0	1	0
.04	.039	.038	.037	.036	.035	.034	.033	.03	.01

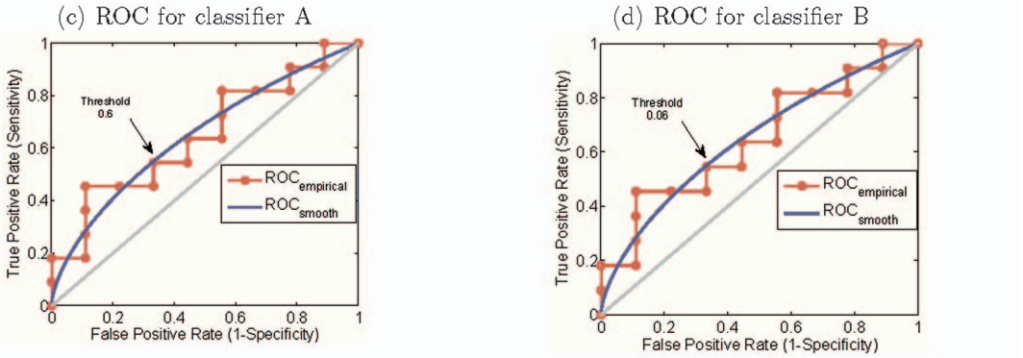


Figure 1. An example of outputs for two probabilistic classifiers and their ROC curves, which do not evaluate calibration. In (a) and (b), # indicates the observations, y corresponds to the class membership, and $\hat{y}$ represents the probability estimate. In (c) and (d), each red circle corresponds to a threshold value. Note that probabilistic classifier B has the same ROC as probabilistic classifier A, but their calibration differs dramatically: estimates for B are ten times lower than estimates for A.
doi:10.1371/journal.pone.0048823.g001

model. Note that we used X and y to denote the features and class label of an observation because the former represents a vector, while the latter refers to a scalar. In this article, we first investigate relationships between *discrimination* and *calibration*, then we proceed to show why it is beneficial to optimize *discrimination* and *calibration* simultaneously. We developed the Doubly Optimized Calibrated Support Vector Machine (DOC-SVM) algorithm that combines the optimization of *discrimination* and *calibration* in a way that can be controlled by the users. We evaluated our approach using real-world data and demonstrated performance advantages when we compared to widely used classification algorithms, i.e., Logistic Regression [13] and Support Vector Machine [14,15].

Methods

Ethics Statement

We use two sets of breast cancer gene expression data with corresponding clinical data downloaded collected from the NCBI Gene Expression Omnibus,i.e., WANG (GSE2034) and SOTIR-ITOU (GSE2990), as well as another breast cancer gene expression data from Chanrion's group [16] in studying the occurrence of relapse as a response to tamoxifen. Because all these data are publicly available, we do not need IRB approval to use them.

Preliminaries

We first review *discrimination* and *calibration* before introducing details of our methodology.

The Area Under ROC Curve (AUC) is often used as a *discrimination* measure of the quality of a probabilistic classifier, e.g., a random classifier like a coin toss has an AUC of 0.5; a perfect classifier has an AUC of 1. Every point on a ROC curve corresponds to a threshold that determines a unique pair of True Positive Rate ($\mathrm{TPR}=\frac{TP}{P}$) and False Positive Rate ($\mathrm{FPR}=\frac{FP}{N}$), where TP, FP, P and N correspond to the number of true positive, false negative, positive, and negative observations, respectively. The AUC can be defined as the integral of TPR (also called *sensitivity*) over FPR (corresponds to *1-specificity*):

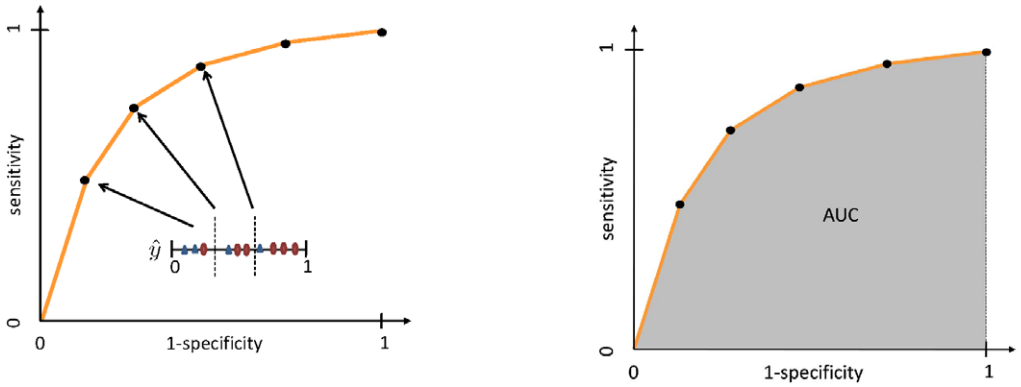
$$\begin{aligned} AUC &= \int_0^1 \frac{TP}{P} d\frac{FP}{N} \\ &= \frac{1}{PN}\int_0^n TP\,dFP \\ &= \frac{1}{PN}\sum_{X\in\{+\}}\sum_{O\in\{-\}}(p(X)\geq p(O)), \end{aligned} \tag{1}$$

where $p(X)$ and $p(O)$ correspond to the estimates for a positive observation X and a negative observation O, respectively. Note that P and N are the counts of positive and negative observations. The last line of Equation (1) corresponds to the result showed in [17] that AUCs can be seen as the total number of concordant pairs out of all positive and negative pairs, which is also known as the c-index [18]. For example, if all positive observations rank higher or the same as the negative observations, the AUC becomes 1; on the other hand, if none of the positive observations rank higher than any of the negative observations, the AUC value is 0. Figure 2 illustrates the relationships between ROC, AUC and its calculation. More details on parametric calculation of AUCs, please refer to [19].

Calibration is a degree of agreement between predicted probability with actual risk, which can be used to evaluate whether a probabilistic classifier is reliable (i.e., faithful representative of the true probability). A probabilistic classifier assigns a probability $\hat{y}_i$ to each observation i. In a perfectly calibrated classifier, the estimated prediction $\hat{y}_i$ is equivalent to the percentage of positive events out of the population that receives this score (e.g., for a group of patients who receive a score of 0.25, one fourth will be positive for the outcome of interest, such as breast cancer). When there are few observations with the same probability, observations

(a) ROC. The dotted lines represent true positive rates and false negative rates determined by thresholds.

(b) The area under the ROC curve corresponds to the AUC.



(c) Concordant pair.

(d) Discordant pair.

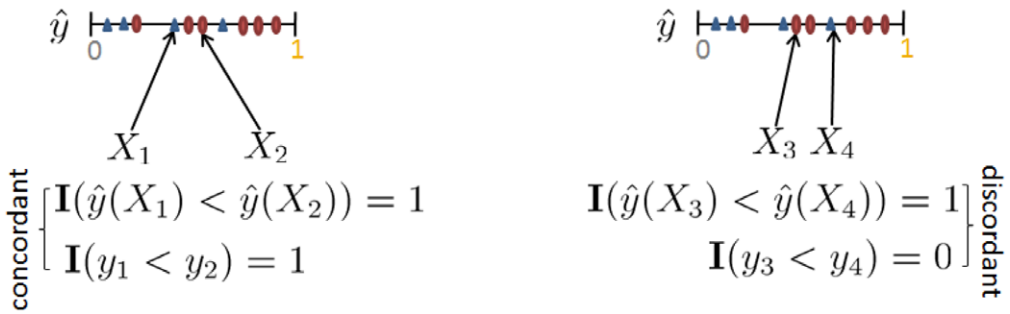


Figure 2. ROC, AUC and its calculation. The horizontal line shows sorted probabilistic estimates on "scores" $\hat{y}$s. In (a) and (b), we show the ROC and the AUC for a classifier built from an artificial dataset. In (c) and (d), we show concordant and discordant pairs, where concordant means that an estimate for a positive observation is higher than an estimate for a negative one. The AUC can be interpreted in the same way as the c-index: the proportion of concordant pairs. Note that X_i corresponds to an observation, $\hat{y}(X_i)$ represents its predicted score, and y_i represents its observed class label, i.e., the gold standard. AUC is calculated as the fraction of concordant pairs out of a total number of instance pairs where an element is positive and the other is negative. Note that $\mathbf{I}(\cdot)$ is the indicator function.
doi:10.1371/journal.pone.0048823.g002

with similar probabilities are grouped by partitioning the range of predictions into groups (or bins). For instance, observations that were assigned estimates between 0.2 and 0.3 may be grouped into the same bin. To estimate the unknown true probabilities for many real problems, it is common to divide the prediction space into ten bins. Observations with predicted scores between 0 and 0.1 fall in the first bin, between 0.1 and 0.2 in the second bin, etc. For each bin, the mean of predicted scores is plotted against the fraction of positive observations. If the model is well calibrated, the points will fall near the diagonal line, as indicated in Figure 3(a).

Calibration can also be measured by goodness-of-fit test statistic, a discrepancy measure between the observed value form the data and the the expected values under the model under consideration.. A widely used goodness-of-fit test in logistic regression is the Hosmer-Lemeshow test (HL-test) [20]. Although the HL-test has important limitations, few practical alternatives have been proposed. In addition, most of these alternatives are model-specific *calibration* measurements, which make them unattractive for evaluating probabilistic outputs (i.e., "scores") across different models. For practical purpose, we use the HL-test as a measure of calibration in this article. The HL-test statistic can be written as $H = \sum_{g=1}^{G} [\frac{(o_g - \pi_g)^2}{\pi_g(1-\pi_g/N_g)}]$, where o_g, N_g and π_g correspond to observed positive events, number of total observations, and the sum of predicted scores for the g^{th} risk bin, respectively. H_H is the Hosmer-Lemeshow H test statistic if a bin is defined by equal-length subgroups of fitted risk predictions, e.g., [0–0.1], [0.1–0.2], …, [0.9–1]; H_C is the Hosmer-Lemeshow C test statistic with an equal number of predicted scores in each group, e.g., m elements in group 1, m elements in group 2, …, m elements in the group G. Usually, elements are divided into ten groups ($G = 10$), and the distribution of the statistics H is approximated by a χ^2 with $G-2$ degrees of freedom, where G indicates the number of groups. Figure 3 illustrates a reliability diagram and two types of the HL-test. Note that in Figure 3(a), the small point-to-line distances roughly indicate that the model is reasonably *calibrated*, and it is not consistently optimistic or pessimistic.

Discrimination-Calibration Tradeoff. Ideally, we want a model with good *discrimination* (i.e., $AUC \approx 1$) and good *calibration* (i.e., $H_C \approx 0$). A perfect model occurs only when predictions are dichotomous (0 or 1) and predictions match observed class labels exactly. There are few conditions in which such black and white cases exit in the real-world. Even under such cases, the result might indicate the model overfits the training data. Figure 4(a) illustrates the situation of perfect *discrimination* and *calibration* in a training set. This usually does not guarantee the same behavior in the test set. Therefore, a realistic concern is whether *calibration* could be harmful to *discrimination*, and vice versa. In other words,

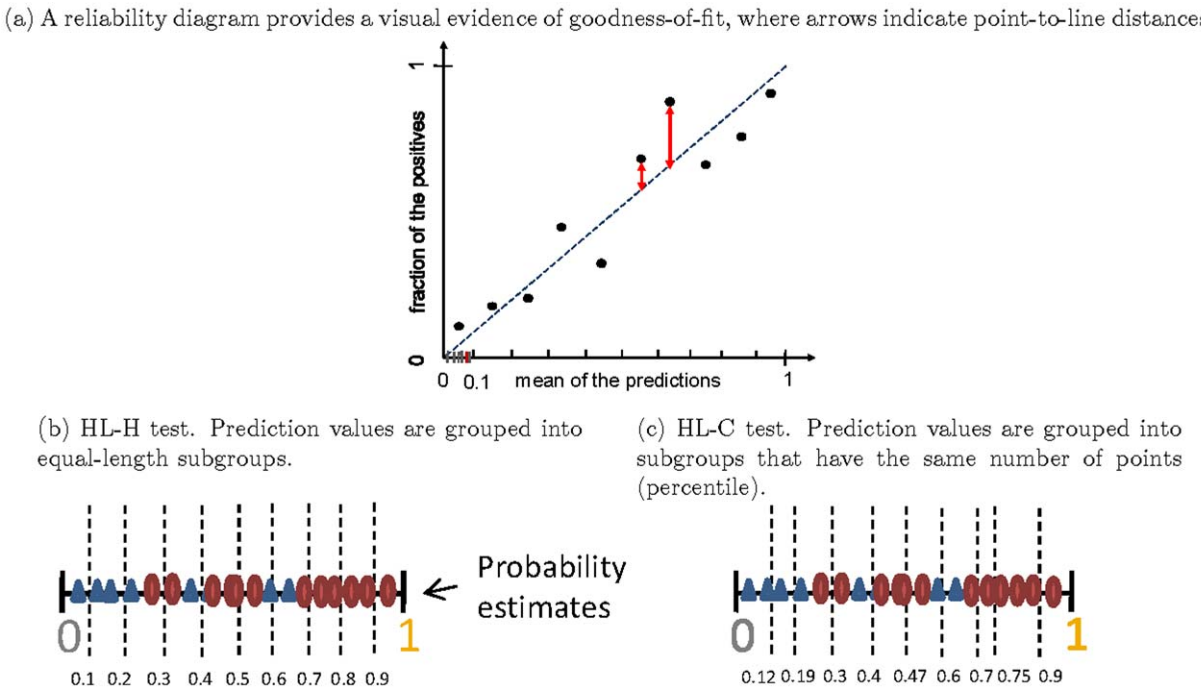


Figure 3. Reliability diagrams and two types of HL-test. In (a), (b), and (c), we visually illustrate the reliability diagram, and groupings used for the HL-H test and the HL-C test, respectively.
doi:10.1371/journal.pone.0048823.g003

suppose we construct a *calibrated* version of some classifier whose predictions are not dichotomous, could this increase the ranking error and hence decrease discrimination?

Figure 4(a, b, c, d) illustrates the relationship between *calibration* and *discrimination* with individual predictions $\hat{y}_i$, derived from a set of probabilistic models. Each subfigure illustrates ten models sampled at AUCs close to a given value (0.5, 0.8, and 0.95). In each column, the upper row represents the ROC plot, and the bottom row corresponds to the reliability diagram (i.e., *calibration plot*). In the ROC plot, grey curves denote empirical ROCs and the bold blue curve represents the averaged smooth ROC. In every *calibration* plot, we show the boxplot and histogram of observed event rates at predicted event rate intervals from 0.1 to 1. Note

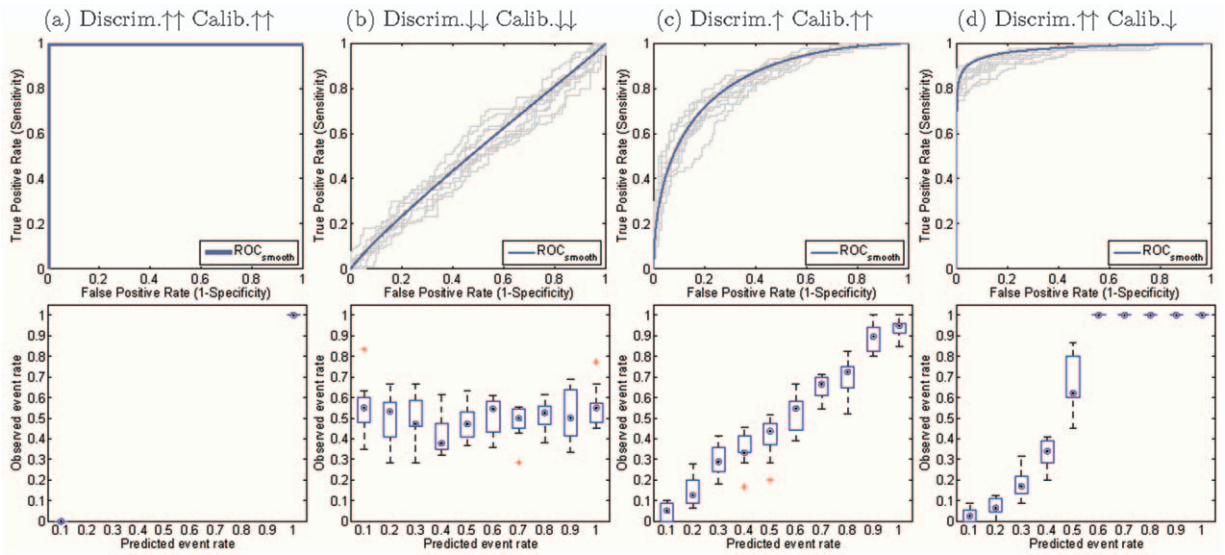


Figure 4. *Discrimination* plots (ROC curves) and *Calibration* plots for simulated models. (a) Perfect *discrimination* (i.e., AUC = 1) requires a classifier with perfect dichotomous predictions, which in the calibration plot has only one point (0,0) for negative observations and one point (1,1) for positive observations. (b) Poor *discrimination* (i.e., AUC = 0.53 ± 0.02) and poor *calibration* (i.e., $H_C = 251.27 \pm 65.2$, $p < 1e-10$). (c) Good *discrimination* (i.e., AUC = 0.83 ± 0.03) and excellent *calibration* (i.e., $H_C = 10.02 \pm 4.42$, $p = 0.26 \pm 0.82$). (d) Excellent *discrimination* (i.e., AUC = 0.96 ± 0.01) and mediocre *calibration* (i.e., $H_C = 34.46 \pm 2.77$, $p = 0 \pm 0.95$). Note that a HL statistic smaller than 13.36 indicates that the model fits well at the significance level of 0.1.
doi:10.1371/journal.pone.0048823.g004

that a good *calibration* would be represented by boxplots that are roughly aligned with the 45 degree line..

The order of (b, d and c) shows that improvement in *calibration* on non-dichotomous predictions may lead to better *discrimination*, but further improvements in *calibration* might result in worse *discrimination*. The reason is that the perfect *calibration* for non-dichotomous predictions has to introduce discordant pairs (indicated by the red arrows in Figure 5(a)) to produce a match between the mean of predictions and the fraction of positive observations within each sub-group. Therefore, the model is prevented from being perfectly *discriminative*. This conclusion is concordant with the result of Diamond [21], who stated that the AUC of a perfectly non-trivially *calibrated* model (constructed from a unique, complementary pair of triangular beta distributions) cannot be over 0.83. Similarly, enforcing *discrimination* might hurt *calibration* as well. Figure 5(b) illustrates this situation using artificial data. Clearly, there is a tradeoff between *calibration* and *discrimination*, and we will explore it in more detail in the next section.

Joint Optimization Framework. We will show that *discrimination* and *calibration* are associated aspects of a well-fittedprobabilistic model, and therefore, they should be jointly optimized for better performance. We start introducing this global learning framework by reviewing the Brier score decomposition.

Brier Score Decomposition: The expectation of squared-losses between y_i and $\hat{y}_i$ is also called *Brier score* [22]

$$\ell_{squared} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

where n is the number of observations. Some algebraic manipulation leads to the following decomposition.

Lemma 1. *The Brier score can be expressed as*

$$\ell_{squared} = \sum_{s}^{|S|}\alpha(s)(s-\beta(s))^2 + \sum_{s}^{|S|}\alpha(s)\beta(s)(1-\beta(s))$$

where $\alpha(s) = \frac{1}{n}|I_s|$, $\beta(s) = \frac{1}{n\alpha(s)}\sum_{i\in I_s} y_i$, and n is the total number of observations, and s is a particular prediction value or score [23].

Proof. To prove that the Brier score can be decomposed into two components, we cluster predictions with the same estimated score s. Thus $\alpha(s)$ is the fraction of times that we predict the score s, and $\beta(s)$ is the fraction of times that the event $y=1$ happens when we predict a score s. Note that I_s indicates a set of instances $\{i\}$ such that $\hat{y}_i = s$, and $|I_s|$ corresponds to the cardinality of the set.

$$\ell_{\text{squared}} = \frac{1}{n}\sum_{i=1}^{n}(y_i^2 - 2\hat{y}_i y_i + \hat{y}_i^2)$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - 2\hat{y}_i y_i + \hat{y}_i^2)$$

$$= \frac{1}{n}\sum_{s}^{|S|}\left(\sum_{j\in I_s} y_j - 2s\sum_{j\in I_s} y_j + \sum_{j\in I_s}\hat{y}_j^2\right)$$

$$= \sum_{s}^{|S|}\left(\alpha(s)\beta(s) - 2s\alpha(s)\beta(s) + \alpha(s)s^2\right)$$

$$= \sum_{s}^{|S|}\left(\alpha(s)s^2 - 2s\alpha(s)\beta(s) + \alpha(s)\beta^2(s) + \alpha(s)\beta(s) - \alpha(s)\beta^2(s)\right)$$

$$= \sum_{s}^{|S|}\alpha(s)(s-\beta(s))^2 + \sum_{s}^{|S|}\alpha(s)\beta(s)(1-\beta(s))$$

There are several versions of Brier score decompositions [24,25,26,27], but for the interest of this article, we will focus on the above two-component decomposition. The first term of Brier score corresponds to *dis-calibration* (**D**) and its minimization encourages $\beta(s) = s$, which is the exact condition required for

(a) Enforcing too much *calibration* may hurt *discrimination*. (b) Enforcing too much *discrimination* may hurt *calibration*.

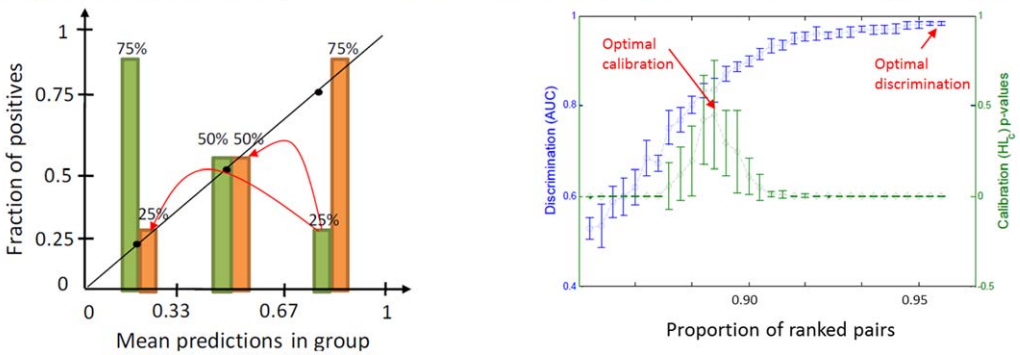


Figure 5. Tradeoffs between *calibration* and *discrimination*. (a) Perfect *calibration* may harm *discrimination* under a three-group binning. The numbers above each bar indicate the percentage of negative observations (green) and positive observations (orange) in each prediction group (0–0.33, 0.33–0.67, and 0.67–1). Note the small red arrows in the left figure indicate discordant pairs, in which negative observations ranked higher than positive observations. (b) Enforcing discrimination may also hurt calibration. The blue curve and error bars correspond to the AUC while the green curve and error bars represent the p-values for the Hosmer-Lemeshow C test (H_C). Initially, as *discrimination* increases, p-value of H_C (*calibration*) increases but it quickly drops after hitting the global maximum. We use red arrows in Figure 5(b) to indicate the location of optimal calibration and discrimination for the simulated data.
doi:10.1371/journal.pone.0048823.g005

well-*calibrated* estimations. The minimization of the second term, called *refinement* [28], encourages $\beta(s)$ to be confident (i.e., close to 0 or 1). The *refinement* term (**R**), which indicates the homogeneity of predicted scores, is closely related to *discrimination*.

The Refinement Term. *Refinement* is a measure of *discrimination* but is often overlooked in favor of its sibling, AUC. Here, we study its properties more closely.

Lemma 2. *We can re-express refinement as*

$$\mathbf{R}=\sum_i \alpha(s_i)\beta(s_i)(1-\beta(s_i))=\frac{1}{N+P}\sum_i N_{s_i}\beta(s_i)(1-\beta(s_i)).$$

Note that $N_{s_i}=|I_{s_i}|$ indicates the number of examples with the predicted score s_i, while N and P correspond to the number of negative and positive examples, respectively.

Lemma 3. *We can re-express the AUC calculated by the trapezoidal method [29] as*

$$\mathbf{A}=\sum_i \frac{1}{2}\frac{N_{s_i}(1-\beta(s_i))}{N}\left(2\sum_{j=1}^{i-1}\frac{N_{s_j}\beta(s_j)}{P}+\frac{N_{s_i}\beta(s_i)}{P}\right). \quad (2)$$

Proof. Each point of the ROC curve has width and height:

$$W(s)=\frac{N_s(1-\beta(s))}{N}$$

$$H(s)=\frac{N_s\beta(s)}{P},$$

thus, the AUC can be approximated by summing over the trapezoidal areas under it:

$$\mathbf{A}=\sum_{i=1}^{|S|}\frac{1}{2}W(s_i)(Total\ height(s_{i-1})+Total\ height(s_i))$$

$$=\sum_i \frac{1}{2}W(s_i)(\sum_j^{i-1}H(s_j)+\sum_j^{i}H(s_j))$$

$$=\sum_i \frac{1}{2}W(s_i)(2\sum_j^{i-1}H(s_j)+H(s_i))$$

$$=\sum_i \frac{1}{2}\frac{N_{s_i}(1-\beta(s_i))}{N}(2\sum_j^{i-1}H(s_j)+H(s_i)).$$

Theorem 4. *AUC is lower bounded by refinement:* $\frac{2NP}{N+P}\mathbf{A}\geq\mathbf{R}$.

Proof. We can reorganize Equation (2) as

$$\mathbf{A}=\frac{1}{2NP}\left(\sum_{i,j<i}2N_{s_i}N_{s_j}\beta(s_i)(1-\beta(s_j))+\sum_i N_{s_i}^2\beta(s_i)(1-\beta(s_i))\right) \quad (3)$$

Because $N_{s_i}\geq 1$,

$$\sum_i N_{s_i}\beta(s_i)(1-\beta(s_i))\leq\sum_i N_{s_i}^2\beta(s_i)(1-\beta(s_i)).$$

Thus, if we multiply $\frac{2NP}{N+P}$ to both sides of Equation (3) and reorganize it,

$$\frac{2NP}{N+P}(\mathbf{A}-\frac{1}{2NP}\sum_{i,j<i}2N_{s_i}N_{s_j}\beta(s_i)(1-\beta(s_j)))$$

$$=\frac{1}{N+P}\sum_i N_{s_i}^2\beta(s_i)(1-\beta(s_i))$$

$$\geq\frac{1}{N+P}\sum_i N_{s_i}\beta(s_i)(1-\beta(s_i))$$

$$=\mathbf{R}$$

Since $\frac{1}{2NP}\sum_{i,j<i}2N_{s_i}N_{s_j}\beta(s_i)(1-\beta(s_j))\geq 0$, we showed that $\frac{2NP}{N+P}\mathbf{A}\geq\mathbf{R}$.

Theorem 4 indicates that maximizing *refinement* encourages the maximization of AUC, which is a critical result for combining *calibration* and *discrimination* into a unified framework.

Doubly Optimized Calibrated Support Vector Machine. We developed a novel approach called Doubly Optimized Calibrated Support Vector Machine (DOC-SVM) to jointly optimize *discrimination* and *calibration*. We will quickly review SVM to help understand the notation we use to explain DOC-SVM. Consider a training dataset $\mathbf{D}=\{(X_1,y_1),\ldots,(X_n,y_n)\}\subset\mathbf{X}\times\mathbf{R}$, where $\mathbf{X}$ denotes the space of input patterns (e.g. $\mathbf{X}=\mathbf{R}^d$) and class labels $y_i\in\{-1,+1\}$. Here "+1" indicates a positive case and "−1" indicates a negative case. A Support Vector Machine (SVM) [15] approximates the zero-one loss by maximizing the geometric margin $\|W\|^2$ between two classes of data. The function it optimizes can be written as

$$\min_{W,\xi}\left[\frac{1}{2}W^TW+C\sum_{i=1}^n\xi_i\right] \quad (4)$$

$$s.t.\ y_iW^TX_i\geq 1-\xi_i$$

$$\xi_i\geq 0,\forall i,$$

where ξ_i is the loss for the i-th data point X_i; W are weight parameters; and C is a penalty parameter to weight the loss. We can reorganize Equation (3) by absorbing the constraints into the objective function

$$\min_W \left[\frac{1}{2}||W||^2 + C\sum_{i=1}^{n} \max(1-y_i W^T X_i, 0)\right]. \quad (5)$$

The first term $\frac{1}{2}||W||^2$ is responsible for the model's complexity. The second term $\max(1-y_i W^T X_i, 0)$, known as the hinge loss ℓ_{hinge}, penalizes the model for mis-classifications. SVM expects label "1" cases to be $f(X) = W^T X > 0$ and label "-1" cases to be $f(X) < 0$. The final output of this optimization is a vector of weight parameters, W, which forms a decision boundary that maximizes the margin between positive and negative cases.

As the hinge loss function only deals with decision boundary, SVM suffices in tasks where the mission is to provide good *calibration* besides *discrimination*. Some researchers proposed ad-hoc post-processing steps like Platt scaling [30] or Isotonic Regression [31] to rectify its output. Our idea is to introduce a second term, the squared loss, to be optimized concurrently with the hinge loss of the original SVM. As we discussed before, the squared loss (Brier Score) can be decomposed into *calibration* and *refinement* components. The major challenge for explicitly controlling the joint optimization is to integrate the *refinement* component with the *hinge loss* component to get a unified *discrimination* component. As we already know there is a relationship between *refinement* and AUC, the challenge boils down to identifying the relationship between the *hinge loss* and the AUC. There are some related empirical studies by Steck and Wang showing that the minimization of the hinge loss leads to the maximization of the AUC [32,33] but we are the first to give a formal proof. Our proof is an extension of Lemma 3.1 in [34].

Theorem 5. *Rank loss (i.e., one minus the Area Under the ROC curve) is bounded by the hinge loss as* $1-AUC \le \frac{1}{\min(p,1-p)} * \ell_{hinge}$, *where* p *is the probability of the positive class.*

Proof. Given a classifier c and n observed events $(X_i, y_i)_{i=1}^n$, we can build the confusion matrix in Table 1, where TP, FP, FN, and TN denote the counts of true positive, false positive, false negative, and true negative instances, respectively.

The number of maximum discordant pairs Γ is bounded by

$$\Gamma \le TP*FP + FP*FN + FN*TN$$

$$= TP*FP + FP*FN + FN*TN + FP*FN - FP*FN$$

$$= FP*(TP+FN) + FN*(TN+FP) - FP*FN.$$

Dividing both sides by $(TP+FN)*(TN+FP)$, we get

$$\frac{\Gamma}{(TP+FN)*(TN+FP)} \le \frac{FP}{TN+FP} + \frac{FN}{TP+FN} - \frac{FP*FN}{(TP+FN)*(TN+FP)}.$$

We can normalize TP, FP, FN, TN by the total number of records to get TP', FP', FN', TN' and their replacement of the formers to the above equation will not change the inequality. We can simplify the equation to get

$$1-AUC(c) \le \frac{\pi}{1-p} + \frac{\eta}{p} - \frac{\eta\pi}{p(1-p)}$$

where $1-AUC = \frac{\Gamma}{(TP'+FN')*(TN'+FP')}$, $FN' = P(c(X) \le 0, y=1)$ (denoted as η), $FP' = P(c(X) > 0, y=0)$ (denoted as π), $TP'+FN' = p$ as the probability of the positive class, and $TN'+FP' = 1-p$ as the probability of the negative class. Therefore, as in Theorem 3.1 of Kotlowski [34],

$$1-AUC(c) \le \frac{\pi}{1-p} + \frac{\eta}{p}$$

$$\le \frac{\pi+\eta}{\min(p,1-p)}$$

$$= \frac{\ell_{0/1}(c)}{\min(p,1-p)},$$

where $\ell_{0/1}(c) = \sum_i (\ell_{0/1}(c(X_i), y_i)) = \sum_i (I(c(X_i) y_i \le 0))$ indicates the zero-one loss, y_i and $c(X_i)$ are the class label and the prediction score of the i-th element, and $I(\cdot)$ is an indicator function. Since the hinge loss function $\ell_{hinge}(c)$ upper bounds the zero-one loss $\ell_{0/1}(c)$ for an arbitrary classifier c (i.e., $\ell_{0/1}(c) \le \ell_{hinge}(c)$), we proved that $\ell_{rank} = 1-AUC \le \frac{1}{\min(p,1-p)} * \ell_{hinge}$.

Although it provides a loose bound, Theorem 5 indicates that minimizing the hinge loss function leads to AUC maximization because $\ell_{hinge}(c) \to 0$ implies $AUC \to 1$. The following objective function optimizes the Doubly Optimized Calibrated Support Vector Machine (DOC-SVM),

$$\min_{\mathbf{w},\xi} \left[\frac{1}{2} W^T W + c_1 \sum_{i=1}^{N} \xi_i + c_2 \sum_{i=1}^{N} (W^T X_i - y_i)^2\right] \quad (6)$$

$$s.t. y_i W^T X_i \ge 1 - \xi_i$$

$$\xi_i \ge 0, \forall i,$$

where ξ_i is the loss for the i-th data point X_i; W is the weight parameter; c_1 and c_2 are the penalty parameters for the hinge loss the squared loss, respectively. DOC-SVM optimizes *discrimination* and *calibration* in a joint manner. Let us denote the hinge loss as ℓ_{hinge}, the squared loss as $\ell_{squared}$, *refinement* as **R**, *dis-calibration* as **D**, and AUC as **A**. Holding the regularization term $\frac{1}{2} W^T W$ as a

Table 1. Confusion matrix of a classifier c based on the gold standard of class labels.

		"Gold standard"	
		Positive	**Negative**
Predictions	Predicted positive	True Positive (FP)	**False Positive (FP)**
	Predicted negative	**False Negative (FN)**	True Negative (TN)

doi:10.1371/journal.pone.0048823.t001

Table 2. Details of the training and test datasets in our first experiment.

	#ATTR	TRAINING SET SIZE	TEST SET SIZE	%POS
GSE2034	15	125	84	54%
GSE2990	15	54	36	67%

doi:10.1371/journal.pone.0048823.t002

constant, Equation (6) concurrently optimizes *discrimination* and *calibration*, and it allows the explicit adjustment of the tradeoff between the two giving,

$$\min(c_1\ell_{hinge}+c_2\ell_{squared})$$

$$: -b_1\max(\mathbf{A})+b_2\min(\mathbf{R}+\mathbf{D})$$

$$: -b_1\max(\mathbf{A})+b_2\min(\mathbf{R})+b_2\min(\mathbf{D})$$

$$: -b_1\max(\mathbf{A})-kb_2\max(\mathbf{A})+b_2\min(\mathbf{D})$$

$$: -(b_1+kb_2)\min(\mathbf{A})+b_2\min(\mathbf{D}),$$

where b_1,b_2 are weight parameters for different loss functions and k is a constant factor. As $\mathbf{A}$ is lower bounded by a factor of $\mathbf{R}$, therefore, the minimization of $\mathbf{A}$ enforces the minimization of $\mathbf{R}$.

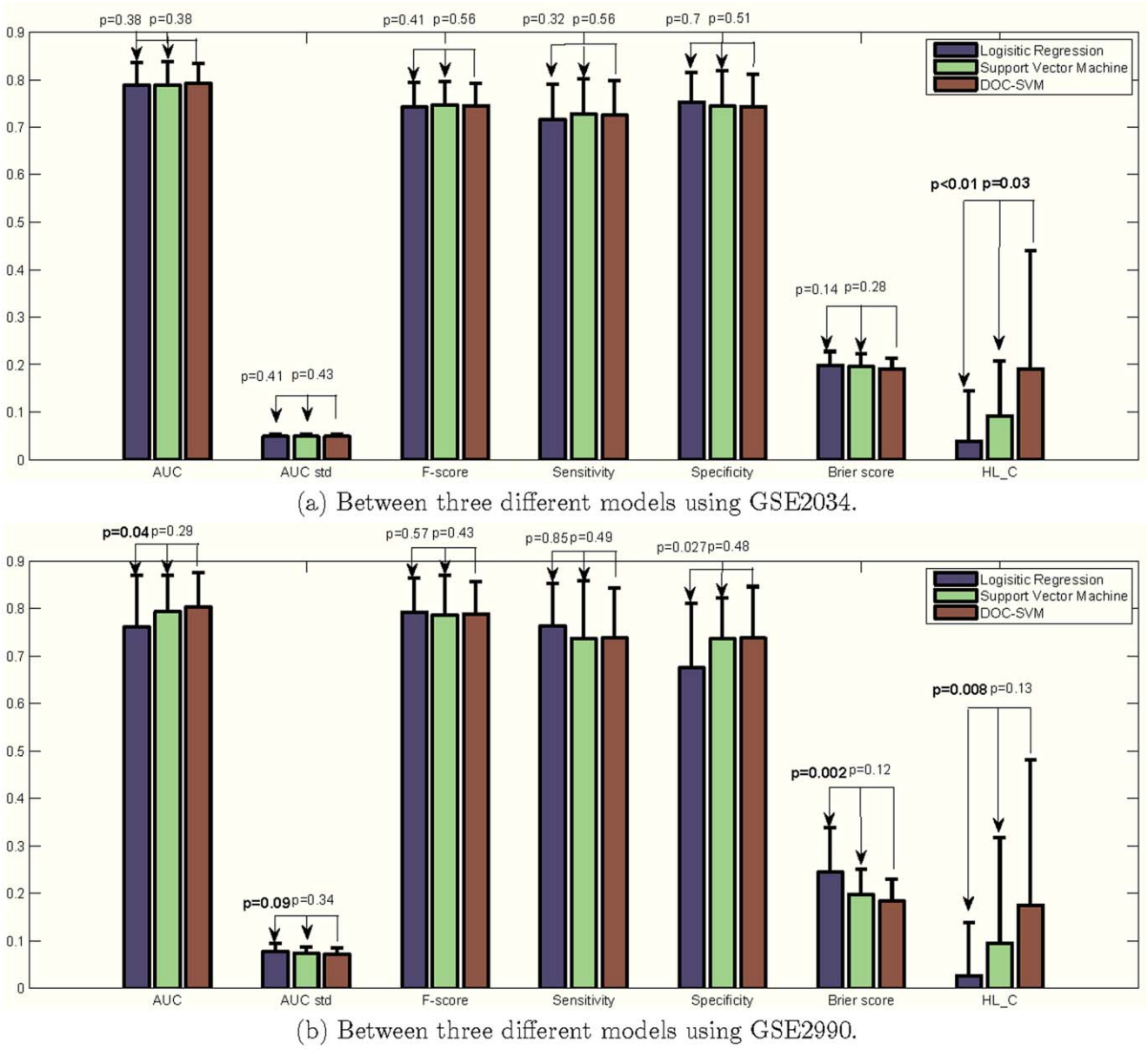


Figure 6. Performance comparison between using GSE2034 and GSE2990.
doi:10.1371/journal.pone.0048823.g006

Table 3. Details of the training and test datasets in our second experiment.

	#ATTR	DATASET SIZE	%POS
Training	36	132	65%
Test	36	23	74%

doi:10.1371/journal.pone.0048823.t003

The higher b_2 is, the less *discriminative* and the more *calibrated* the classifier is, and vice versa.

Experiments

We evaluated the efficacy of DOC-SVM using three real-world datasets. We examined the *calibration* and *discrimination* of the LR, SVM, and DOC-SVM. To allow for a fair comparison, we applied Platt's method to transform SVM's outputs into probabilities [30]. We adopted ten-fold cross validation [35] to pick the best parameters (i.e., c_1, c_2 for DOC-SVM and c for SVM) for each model. Specially, we used the following metrics for evaluation: AUC, AUC standard deviation, F-score, Sensitivity, Specificity, Brier Score, and the p-value of the HL-C test, which are all among the most commonly used in statistical model comparisons. The null hypothesis in our HL-C test is that the data are generated by the model fitted by the researcher. If test statistic is less or equal to 0.1, we reject the null hypothesis that there is no difference between the observed and model-predicted values, which implies that the model's estimates do not fit the data well (i.e., the calibration is poor). Otherwise, if the test statistic is greater than 0.1, as expected for well-fitting models, we fail to reject the null hypothesis.

Our first experiment used breast cancer gene expression data collected from the NCBI Gene Expression Omnibus (GEO). Two individual datasets were downloaded, and were previously studied by Wang et al. (GSE2034) [36] and Sotiriou et al. (GSE2990) [37]. Table 2 summarized the data sets. To make our data comparable to the previous studies, we followed the criteria outlined by Osl et al. [38] to select patients who did not receive any treatment and had negative lymph node status. Among these pre-selected candidates, only patients with extreme outcomes, either poor outcomes (recurrence or metastasis within five years) or very good outcomes (neither recurrence nor metastasis within eight years) were selected. The number of observations after filtering was: 209 for GSE2034 (114 good/95 poor) and 90 for GSE2990 (60 good/30 poor). All of these data have a feature size of 247,965, which corresponds to the gene expression results obtained from certain micro-array experiments. Both datasets have been preprocessed to keep only the top 15 features by T-test, as previously described [38].

Figure 6(a) and Figure 6(b) illustrate a number of comparisons between LR, SVM, and DOC-SVM using the GSE2034 and GSE2990 datasets over 30 random splits. In both experiments, DOC-SVM showed higher AUCs when compared to other models under one-tailed paired t-tests using $p = 0.1$ as the threshold. Although the improvements to SVM are small (GSE2034: $p = 0.38$, GSE2990: $p = 0.29$), DOC-SVM had significantly higher AUCs compared to LR in GSE2990 ($p = 0.04$). Besides discrimination, DOC-SVM demonstrated better calibration in terms of HL-C test. In the experiment, DOC-SVM had significantly higher p-values than the LR model (GSE2034: $p<0.01$, GSE2990: $p = 0.008$) using a one-tailed paired t-test. An improvement to SVM was significant for GSE2034 ($p = 0.03$) but not for GSE2990 ($p = 0.13$). We also conducted one-tailed paired t-tests to evaluate if DOC-SVM has smaller Brier scores when compared to LR and SVM. The results were similar to what we already observed in discrimination and calibration: the Brier scores were smaller than those of LR (GSE2034: $p = 0.14$, GSE2990: $p = 0.002$) and SVM (GSE2034: $p = 0.28$, GSE2990: $p = 0.12$), but not all improvements were significant. In no instances DOC-SVM performed significantly worse than SVM and LR.

Our second experiment used another breast cancer gene expression data, in which Chanrion and his colleagues predicted the occurrence of relapse as a response to tamoxifen [16]. We followed their experimental design, and conducted log2-transformation and median-centering per sample on the measurement values. To ensure consistency, we selected 36 genes present in their

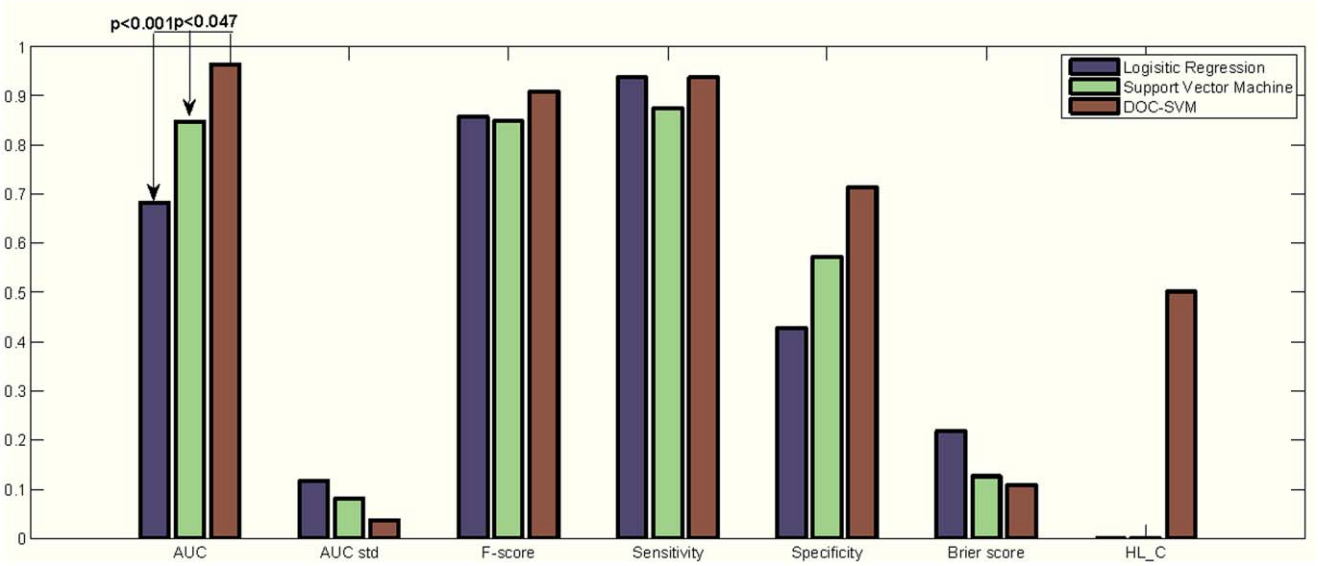


Figure 7. Performance comparisons between three different models using breast cancer datasets.
doi:10.1371/journal.pone.0048823.g007

study and applied nearest shrunken centroid classification method [39]. Note that we carefully split the 155 observations into training and test sets to match what has been reported in that study. Data sets used are shown in Table 3.

The evaluation of this experiment does not involve random split as the training and test datasets were predetermined [16]. Figure 7 shows indices for all three models.

DOC-SVM demonstrated better discrimination performance on the test data (AUC = 0.964), which was significantly higher than the AUCs of SVM (0.848) and LR (0.683). Note that for the comparison of a pair of AUCs, we used a z-test reviewed in Lasko et al. [40]. DOC-SVM also had the lowest Brier score among the three models. In addition, it was the only model that had a good fit, with a HL-C test p-value equals 0.5, whereas p-values of the other models were smaller than 0.0001.

In summary, DOC-SVM showed superior performance in all these real-world datasets. The performance improvements were observed for both *discrimination* and *calibration*, which indicates that DOC-SVM may have better generalization ability compared to LR and SVM due to the joint consideration of both factors. Although these experiments are limited by the small sizes of datasets, their outputs verified our theoretical results and served to demonstrate the advantage of the proposed joint optimization framework.

Conclusions

We explored the properties of *discrimination* and *calibration*, and uncovered an important tradeoff between them, expressed in terms of AUC, a popular measure of *discrimination*. Our investigation also indicated that a supervised probabilistic model can be improved when both *discrimination* and *calibration* are considered in a joint manner. We developed a Doubly Optimized Calibrated Support Vector Machine Model (DOC-SVM) to minimize the squared loss concurrently with the hinge loss to account for both aspects of *discrimination* and *calibration*. Experimental results from using real-world breast cancer datasets indicate that the DOC-SVM can potentially outperform Logistic Regression and Support Vector Machine. Further studies are needed to investigate strategies to tune weights for *discrimination* and *calibration* depending on the learning problem.

Supporting Information

Appendix S1 subgradient descent optimization for SVM.
(DOCX)

Appendix S2 parameter tuning using ten-fold cross validation.
(DOCX)

Appendix S3 additional experiments.
(DOCX)

Appendix S4 DOC-SVM Matlab code.
(DOCX)

Acknowledgments

We thank Dr. Michele Day for editing this manuscript. We thank Dr. Melanie Osl and Mr. Zhanglong Ji for the helpful discussion.

Author Contributions

Conceived and designed the experiments: XJ LO AM. Performed the experiments: XJ JK. Analyzed the data: XJ LO JK. Contributed reagents/materials/analysis tools: SW. Wrote the paper: XJ LO JK. Contributed to Methodology Development: AM.

References

1. Inza I, Calvo B, Armananzas R, Bengoetxea E, Larranaga P, et al. (2010) Machine learning: an indispensable tool in bioinformatics. Methods in Molecular Biology 593: 25–48.
2. Chi YY, Zhou XH (2008) The need for reorientation toward cost-effective prediction: comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by Pencina et al. Statistics in Medicine 27: 182–184.
3. Jiang X, Osl M, Kim J, Ohno-Machado L (2011) Calibrating predictive model estimates to support personalized medicine. Journal of the American Medical Informatics Association 19: 263–274.
4. James AH, Barbara JM (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143.
5. Sarkar IN, Butte AJ, Lussier YA, Tarczy-Hornoch P, Ohno-Machado L (2011) Translational bioinformatics: linking knowledge across biological and clinical realms. Journal of the American Medical Informatics Association 18: 354–357.
6. Wei W, Visweswaran S, Cooper GF (2011) The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data. Journal of the American Medical Informatics Association 18: 370–375.
7. Ayer T, Alagoz O, Chhatwal J, Shavlik JW, Kahn Jr CE, et al. (2010) Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration. Cancer 116: 3310–3321.
8. Jiang X, Boxwala A, El-Kareh R, Kim J, Ohno-Machado L (2012) A patient-Driven Adaptive Prediction Technique (ADAPT) to Improve Personalized Risk Estimation for Clinical Decision Support. Journal of American Medical Informatics Association 19: e137–e144.
9. Lipson J, Bernhardt J, Block U, W.R F, Hofmeister R, et al. (2009) Requirements for calibration in noninvasive glucose monitoring by Raman spectroscopy. Journal of Diabetes Science and Technology 3: 233–241.
10. Sarin P, Urman RD, Ohno-Machado L (2012) An improved model for predicting postoperative nausea and vomiting in ambulatory surgery patients using physician-modifiable risk factors. Journal of the American Medical Informatics Association [Eput ahead of print].
11. Wu Y, Jiang X, Kim J, Ohno-Machado L (2012) Grid LOgistic REgression (GLORE): Building Shared Models Without Sharing Data. Journal of American Medical Informatics Association (Epub ahead of print).
12. Ediger MN, Olson BP, Maynard JD (2009) Personalized medicine for diabetes: Noninvasive optical screening for diabetes. Journal of Diabetes Science and Technology 3: 776–780.
13. Hosmer DW, Lemeshow S (2000) Applied logistic regression. New York: Wiley-Interscience.
14. Vapnik NV (2000) The nature of statistical learning theory. New York: Springer-Verlag.
15. Vapnik NV (1999) An overview of statistical learning theory. IEEE Transactions on Neural Network 10: 988–999.
16. Chanrion M, Negre V, Fontaine H, Salvetat N, Bibeau F, et al. (2008) A gene expression signature that can predict the recurrence of tamoxifen-treated primary breast cancer. Clinical Cancer Research 14: 1744–1752.
17. Bamber D (1975) The area above the ordinal dominance graph and the area below the receiver operating graph. Journal of Mathematical Psychology 12: 385–415.
18. Harrell F, Califf R, Pryor D, Lee K (1982) Evaluating the yield of medical tests. JAMA 247: 2543–2546.
19. Zou KH, Liu AI, Bandos AI, Ohno-Machado L, Rockette HE (2011) Statistical evaluation of diagnostic performance: topics in ROC analysis. Boca Raton, FL: Chapman & Hall/CRC Biostatistics Series.
20. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S (1997) A comparison of goodness-of-fit tests for the Logistic Regression model. Statistics in Medicine 16: 965–980.
21. Diamond GA (1992) What price perfection? Calibration and discrimination of clinical prediction models Journal of Clinical Epidemiology 45: 85–89.
22. Brier GW (1950) Verification of forecasts expressed in terms of probability. Monthly Weather Review 78: 1–3.
23. Sanders F (1963) On subjective probability forecasting. Journal of Applied Meteorology 2: 191–201.
24. Murphy AH (1986) A new decomposition of the Brier score: Formulation and interpretation. Monthly Weather Review 114: 2671–2673.
25. Murphy AH (1973) A new vector partition of the probability score. Journal of Applied Meteorology 12: 595–600.
26. Murphy AH (1971) A note on the ranked probability score. Journal of Applied Meteorology 10: 155.
27. Yaniv I, Yates JF, Smith JK (1991) Measures of discrimination skill in probabilistic judgment. Psychological Bulletin 110: 611–617.
28. DeGroot MH, Fienberg SE (1981) Assessing probability assessors: calibration and refinement. Technical Report. Pittsburgh, PA: Carnegie Mellon University.

29. Purves RD (1992) Optimum numerical integration methods for estimation of area-under-the-curve (AUC) and area-under-the-moment-curve (AUMC). Journal of Pharmacokinetics and Pharmacodynamics 20: 211–226.
30. Platt JC (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers: 61–74.
31. Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates; 2002. pp. 694–699.
32. Steck H (2007) Hinge rank loss and the area under the ROC curve. European Conference on Machine Learning (ECML). Warsaw, Poland. pp. 347–358.
33. Wang Z (2011) HingeBoost: ROC-based boost for classification and variable selection. International Journal of Biostatistics 7: 1–30.
34. Kotlowski W, Dembczynski K, Hüllermeier E (2011) Bipartite Ranking through Minimization of Univariate Loss; 2011; Bellevue, WA. pp. 1113–1120.
35. Duda RO, Hart PE, Stor DG (2001) Pattern classification. New York: Wiley.
36. Wang Y, Klijn J, Zhang Y, Sieuwerts A, Look M, et al. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365: 671–679.
37. Sotiriou C, Wirapati P, Loi S, Harris A S F (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. Journal of the National Cancer Institute 98: 262–272.
38. Osl M, Dreiseitl S, Kim J, Patel K, Baumgartner C, et al. (2010) Effect of data combination on predictive modeling: a study using gene expression data; 2010; Washington D.C. pp. 567–571.
39. Tibshirani R, Hastie T, Narasimhan B, Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences of the United states of America 99: 6567–6572.
40. Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L (2005) The use of receiver operating characteristic curves in biomedical informatics. Journal of Biomedical Informatics 38: 404–415.