Obtaining Well Calibrated Probabilities Using Bayesian Binning

Mahdi Pakdaman Naeini[1], **Gregory F. Cooper**[1,2], **and Milos Hauskrecht**[1,3]
[1]Intelligent Systems Program, University of Pittsburgh, PA, USA
[2]Department of Biomedical Informatics, University of Pittsburgh, PA, USA
[3]Computer Science Department, University of Pittsburgh, PA, USA

Abstract

Learning probabilistic predictive models that are well calibrated is critical for many prediction and decision-making tasks in artificial intelligence. In this paper we present a new non-parametric calibration method called Bayesian Binning into Quantiles (BBQ) which addresses key limitations of existing calibration methods. The method post processes the output of a binary classification algorithm; thus, it can be readily combined with many existing classification algorithms. The method is computationally tractable, and empirically accurate, as evidenced by the set of experiments reported here on both real and simulated datasets.

Introduction

A rational problem solving agent aims to maximize its utility subject to the existing constraints (Russell and Norvig 1995). To be able to maximize the utility function for many practical prediction and decision-making tasks, it is crucial to develop an accurate probabilistic prediction model from data. Unfortunately, many existing machine learning and data mining models and algorithms are not optimized for obtaining accurate probabilities and the predictions they produce may be miscalibrated. Generally, a set of predictions of a binary outcome is well calibrated if the outcomes predicted to occur with probability p do occur about p fraction of the time, for each probability p that is predicted. This concept can be readily generalized to outcomes with more than two values. Figure 1 shows a hypothetical example of a reliability curve (DeGroot and Fienberg 1983; Niculescu-Mizil and Caruana 2005), which displays the calibration performance of a prediction method. The curve shows, for example, that when the method predicts $Z = 1$ to have probability 0.5, the outcome $Z = 1$ occurs in about 0.57 fraction of the instances (cases). The curve indicates that the method is fairly well calibrated, but it tends to assign probabilities that are too low. In general, perfect calibration corresponds to a straight line from $(0, 0)$ to $(1, 1)$. The closer a calibration curve is to this line, the better calibrated is the associated prediction method.

Producing well-calibrated probabilistic predictions is critical in many areas of science (e.g., determining which experiments to perform), medicine (e.g., deciding which therapy to give a patient), business (e.g., making investment decisions), and others. However, model calibration and the learning of well-calibrated probabilistic models have not been studied in the machine learning literature as extensively as for example discriminative machine learning models that are built to achieve the best possible discrimination among classes of objects. One way to achieve a high level of model calibration is to develop methods for learning probabilistic models that are well-calibrated, *ab initio*. However, this approach would require one to modify the objective function used for learning the model and it may increase the computational cost of the associated optimization task. An alternative approach is to construct well-calibrated models by relying on the existing machine learning methods and by modifying their outputs in a post-processing step to obtain the desired model. This approach is often preferred because of its generality, flexibility, and the fact that it frees the designer of the machine learning model from the need to add additional calibration measures into the objective function used to learn the model. The existing approaches developed for this purpose include histogram binning, Platt scaling, or isotonic regression (Platt 1999; Zadrozny and Elkan 2001; 2002). In all these the post-processing step can be seen as a function that maps the output of a prediction model to probabilities that are intended to be well-calibrated. Figure 1 shows an example of such a mapping.

Existing post-processing calibration methods can be divided into two groups: parametric and nonparametric methods. An example of a parametric method is Platt's method that applies a sigmoidal transformation that maps the output of a predictive model to a calibrated probability output (Platt 1999). The parameters of the sigmoidal transformation function are learned using a maximum likelihood estimation framework. The key limitation of the approach is the (sigmoidal) form of the transformation function, which only rarely fits the true distribution of predictions. The most common non-parametric methods are based either on binning (Zadrozny and Elkan 2001) or isotonic regression (Zadrozny and Elkan 2002). In the histogram binning approach, also known as quantile binning, the raw predictions of a binary classifier are sorted first, and then they are partitioned into B subsets of equal size, called bins. Given a prediction y, the method finds the bin containing that prediction and re-

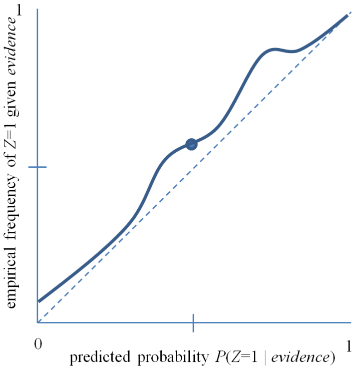


Figure 1: The solid line shows a calibration (reliability) curve for predicting $Z = 1$. The dotted line is the ideal calibration curve.

turns as $\hat{y}$ the fraction of positive outcomes $(Z = 1)$ in the bin. Histogram binning has several limitations, including the need to define the number of bins and the fact that the bins and their associated boundaries remain fixed over all predictions (Zadrozny and Elkan 2002).

The other non-parametric calibration method is based on isotonic regression (Zadrozny and Elkan 2002). This method only requires that the mapping function be isotonic (monotonically increasing) (Niculescu-Mizil and Caruana 2005). A commonly used method for computing the isotonic regression is *pool adjacent violators* (PAV) algorithm (Barlow et al. 1972). The isotonic calibration method based on the (PAV) algorithm can be viewed as a binning algorithm where the position of boundaries and the size of bins are selected according to how well the classifier ranks the examples in the training data (Zadrozny and Elkan 2002). Recently a variation of the isotonic-regression-based calibration method for predicting accurate probabilities with a ranking loss was proposed (Menon et al. 2012). Although isotonic regression based calibration yields good performance in many real data applications, the violation of isotonicity assumption in practice is quite frequent, so the relaxation of the isotonicity constraints may be appropriate.

This paper presents a new binary classifier calibration method called Bayesian Binning into Quantiles (BBQ) that is applied as a post-processing step[1]. The approach can be viewed as a refinement of the histogram-binning calibration method in that it considers multiple different binnings and their combination to yield more robust calibrated predictions. Briefly, by considering only one fixed bin discretization, one may not be able to guess correctly the optimal bin width. In addition, the data may not be distributed equally across the output space after applying the discriminative projection learned in the first step, and various biases in the distribution may occur. Both of these problems can be resolved by considering multiple different binnings and their combination. The experimental results presented below indicate that the proposed method works well in practice.

Methods

BBQ extends the simple histogram-binning calibration method (Zadrozny and Elkan 2001) by considering multiple binning models and their combination. The main challenge here is to decide on how to pick the models and how to combine them. BBQ considers multiple equal-frequency binning models that distribute the data-points in the training set equally across all bins. The different binning models differ in the number of bins they have. We combine them using a Bayesian score derived from the BDeu (Heckerman, Geiger, and Chickering 1995) score used for learning Bayesian network structures. Let y_i and z_i define respectively an uncalibrated classifier prediction and the true class of the i'th instance. Also, let $\mathcal{D}$ define the set of all training instances (y_i, z_i) and let $\mathcal{S}$ be the sorted set of all uncalibrated classifier predictions $\{y_1, y_2 \ldots, y_N\}$, where N is total number of training data. In addition, let Pa denote a partitioning of $\mathcal{S}$ into B equal frequency bins. A binning model M induced by the training set is defined as $M \equiv \{B, Pa, \Theta\}$. Where, B is the number of bins over the set $\mathcal{S}$ and Θ is the set of all the calibration model parameters $\Theta = \{\theta_1, \ldots, \theta_B\}$, which are defined as follows. For a bin b the distribution of the class variable $P(Z = 1|B = b)$ is modeled as a binomial distribution with parameter θ_b. Thus, Θ specifies all the binomial distributions for all the existing bins in Pa. We score a binning model M as follows:

$$Score(M) = P(M) \cdot P(\mathcal{D}|M) \tag{1}$$

The marginal likelihood $P(\mathcal{D}|M)$ in Equation 1 has a closed form solution under the following assumptions (Heckerman, Geiger, and Chickering 1995): (1) All samples are i.i.d and the class distribution $P(Z|B = b)$, which is class distribution for instances located in bin number b, is modeled using a binomial distribution with parameter θ_b, (2) the distributions of the class variable over two different bins are independent of each other, and (3) the prior distribution over binning model parameters θs are modeled using a $Beta$ distribution. We also assume that the parameters of the $Beta$ distribution associated with θ_b are α_b and β_b. We set them to be equal to $\alpha_b = \frac{N'}{B} p_b$ and $\beta_b = \frac{N'}{B}(1 - p_b)$, where N' is the equivalent sample size expressing the strength of our belief in the prior distribution[2] and p_b is the midpoint of the interval defining the b'th bin in the binning model M. Given the above assumptions, the marginal likelihood can be expressed as (Heckerman, Geiger, and Chickering 1995):

$$P(D|M) = \prod_{b=1}^{B} \frac{\Gamma(\frac{N'}{B})}{\Gamma(N_b + \frac{N'}{B})} \frac{\Gamma(m_b + \alpha_b)}{\Gamma(\alpha_b)} \frac{\Gamma(n_b + \beta_b)}{\Gamma(\beta_b)},$$

where Γ is the gamma function and N_b is the total number of training instances located in the b'th bin. Also, n_b and m_b are respectively the number of class *zero* and class *one* instances among all N_b training instances in bin b. The term $P(M)$ in Equation 1 specifies the prior probability of the binning model M. In our experiments we use a uniform prior for modeling $P(M)$. BBQ uses the above Bayesian

[1] An implementation of BBQ method can be found at the following address: https://github.com/pakdaman/calibration.git

[2] We set $N' = 2$ in our experiments.

score to perform model averaging over the space of all possible equal frequency binnings. We could have also used the above Bayesian score to perform the model selection, which in our case would yield a single binning model. However, model averaging is typically superior to model selection (Hoeting et al. 1999). Hence a calibrated prediction in our *BBQ* framework is defined as:

$$P(z=1|y) = \sum_{i=1}^{T} \frac{Score(M_i)}{\sum_{j=1}^{T} Score(M_j)} P(z=1|y, M_i),$$

where T is the total number of binning models considered and $P(z = 1|y, M_i)$ is the probability estimate[3] obtained using the binning model M_i, for the (uncalibrated) classifier output y. To choose models M_i we choose a restricted range of binning models, each defined by a different number of bins. We define the range of possible values of the number of bins as $B \in \{\frac{\sqrt[3]{N}}{C}, \ldots, C\sqrt[3]{N}\}$, where C is a constant that controls the number of binning models ($C = 10$ in our experiments). The choice of the above range is due to some previous results (Klemela 2009; Scott and Nowak 2003) that show that the *fixed bin size* histogram binning classifier is a mini-max rate classifier for Lipschitz Bayes decision boundaries when we set number of bins to $\theta(\sqrt[3]{N})$, where $\theta(.)$ is asymptotically tight bound notation defined as $\theta(g(n)) = \{h(n) : \exists$ positive constants c_1, c_2, n_0 such that $0 \leq c_1 g(n) \leq h(n) \leq c_2 g(n), \forall n \geq n_0\}$. Although the results are valid for histogram classifiers with fixed bin size, our experiments show that both fixed bin size and fixed frequency histogram classifiers behave quite similarly. We conjecture that a histogram classifier with equal frequency binning is also a mini-max rate classifier; this is an interesting open problem that we intend to study in the future.

We may further restrict the number of binning models used in averaging in the application stage. That is, we may start by calculating the Bayesian score for all models in the above range, and select a subset of those that yield a higher Bayesian score afterwards. The number of resulting models can be determined by *a priori* fixing the number of models to be used in averaging or by checking for the sharp drops in the Bayesian scores over all such models. Assume $S_1, S_2, \ldots, S_T$ are the sorted Bayesian scores of histogram models in a decreasing order. We fix a small number $\rho > 0$ ($\rho = 0.001$ in our experiments) and pick the first k_ρ associated binning models as the refined set of models, where $k_\rho = \min\{k : \frac{S_k - S_{k+1}}{\sigma^2} \leq \rho\}$ and σ^2 is the empirical variance of the Bayesian scores.

Calibration Measures

In order to evaluate the calibration capability of a classifier, we use two intuitive statistics that measure calibration relative to the ideal reliability diagram (DeGroot and Fienberg 1983; Niculescu-Mizil and Caruana 2005) (Figure 1 shows an example of a reliability diagram). These measures are called Expected Calibration Error (ECE), and Maximum Calibration Error (MCE). In computing these measures, the predictions are sorted and partitioned into K fixed number of bins ($K = 10$ in our experiments). The predicted value of each test instance falls into one of the bins. The ECE calculates Expected Calibration Error over the bins, and MCE calculates the Maximum Calibration Error among the bins, using empirical estimates as follows:

$$ECE = \sum_{i=1}^{K} P(i) \cdot |o_i - e_i| \quad , \quad MCE = \max_{i=1}^{K} (|o_i - e_i|) ,$$

where o_i is the true fraction of positive instances in bin i, e_i is the mean of the post-calibrated probabilities for the instances in bin i, and $P(i)$ is the empirical probability (fraction) of all instances that fall into bin i. The lower the values of ECE and MCE, the better is the calibration of a model.

Empirical Results

This section describes the set of experiments that we performed to evaluate the performance of the proposed calibration method in comparison to other commonly used calibration methods: histogram binning, Platt's method, and isotonic regression. To evaluate the calibration performance of each method, we ran experiments on both simulated and on real data. For the evaluation of the calibration methods, we used 5 different measures. The first two measures are Accuracy (Acc) and the Area Under the ROC Curve (AUC), which measure discrimination. The three other measures are the *root mean square error* (RMSE), the *expected calibration error* (ECE), and the *maximum calibration error* (MCE), which measure calibration.

Simulated Data

For the simulated data experiments, we used a binary classification dataset in which the outcomes were not linearly separable. The scatter plot of the simulated dataset is shown in Figure 2. The data were divided into 1000 instances for training and calibrating the prediction model, and 1000 instances for testing the models.

To conduct the experiments on simulated datasets, we used two extreme classifiers: *support vector machines* (SVM) with linear and quadratic kernels. The choice of SVM with a linear kernel allows us to see how the calibration methods perform when the classification model makes over simplifying (linear) assumptions. Also, to achieve good discrimination on the data in Figure 2, SVM with a quadratic kernel is an ideal choice. So, the experiment using a quadratic kernel SVM allows us to see how well different calibration methods perform when we use an ideal learner for the classification problem in terms of discrimination.

As seen in Tables 1, BBQ outperforms Platt's method and isotonic regression on the simulation dataset, especially when the linear SVM method is used as the base learner. The poor performance of Platt's method is not surprising given its simplicity, which consists of a parametric model with only two parameters. However, isotonic regression is a non-parametric model that only makes a monotonicity assumption over the output of the base classifier. When we

[3] We actually use smoothing of the counts in the binning models, which is consistent with the Bayesian priors in the scoring function.

	SVM	Hist	Platt	IsoReg	BBQ
AUC	0.50	0.84	0.50	0.65	0.85
ACC	0.48	0.78	0.52	0.64	0.78
RMSE	0.50	0.39	0.50	0.46	0.38
ECE	0.28	0.07	0.28	0.35	0.03
MCE	0.52	0.19	0.54	0.58	0.09

(a) SVM Linear

	SVM	Hist	Platt	IsoReg	BBQ
AUC	1.00	1.00	1.00	1.00	1.00
ACC	0.99	0.99	0.99	0.99	0.99
RMSE	0.21	0.09	0.19	0.08	0.08
ECE	0.14	0.01	0.15	0.00	0.00
MCE	0.35	0.04	0.32	0.03	0.03

(b) SVM Quadratic Kernel

Table 1: Experimental Results on Simulated dataset

use a linear kernel SVM, this assumption is violated because of the non-linearity of data. As a result, isotonic regression performs relatively poorly, in terms of improving the discrimination and calibration capability of the base classifier. The violation of this assumption can happen in real data as well. In order to mitigate this pitfall, Menon et. al (Menon et al. 2012) proposed a new isotonic based calibration method using a combination of optimizing AUC as a ranking loss measure, plus isotonic regression for building an accurate ranking model. However, this is counter to our goal of developing post-processing methods that can be used with any existing classification models. As shown in Table 1b, even if we use an ideal SVM classifier for our linearly non-separable dataset, the proposed method performs as well as an isotonic regression based calibration.

As can be seen in Table 1b, although the SVM based learner performs very well in terms of discrimination based on AUC and ACC measures, it performs poorly in terms of calibration, as measured by RMSE, MCE, and ECE. Moreover, while improving calibration, all of the calibration methods retain the same discrimination performance that was obtained prior to post-processing.

Real Data

In terms of real data, we used 30 different real world binary classification data sets from the UCI and LibSVM repository [4] (Bache and Lichman 2013; Chang and Lin 2011). We used three common classifiers, namely, Logistic Regression (LR), Support Vector Machines (SVM), and Naive Bayes (NB) to evaluate the performance of the proposed calibration method. To evaluate the performance of calibration models, we use the recommended statistical test procedure by

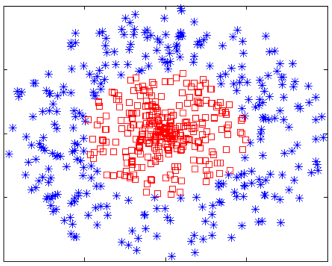
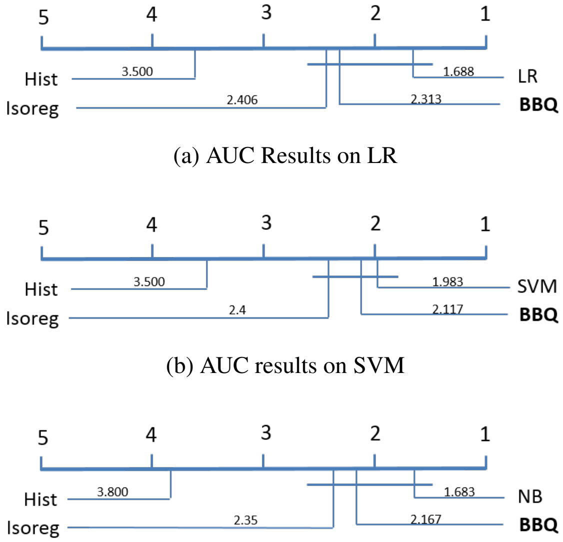


Figure 2: Scatter plot of non-linear separable simulated data



(a) AUC Results on LR

(b) AUC results on SVM

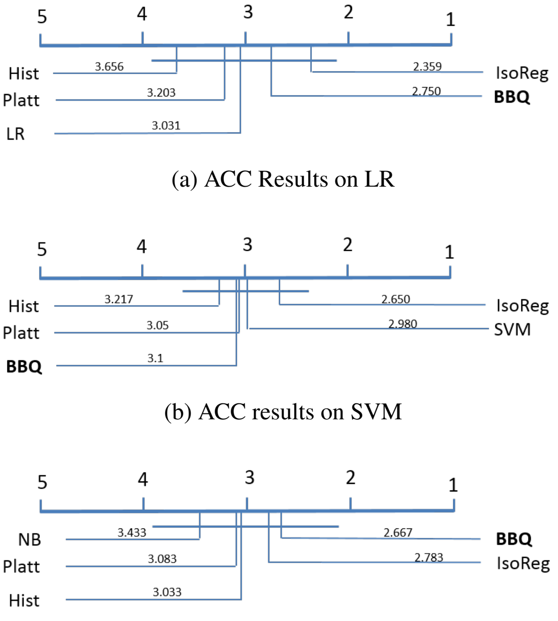
(c) AUC results on NB

Figure 3: Performance of each method in terms of average rank of AUC on the real datasets. All the methods which are not connected to BBQ by the horizontal bar are significantly different from BBQ (using Friedman test followed by Holm's step-down procedure at a 0.05 significance level).

Janez Demsar (Demšar 2006). More specifically, we use the Freidman nonparametric hypothesis testing method (Friedman 1937) followed by Holm's step-down procedure (Holm 1979) to evaluate the performance of BBQ in comparison to the other calibration methods across the 30 real data sets. Next, we briefly describe the test procedure; more detailed information can be found in (Demšar 2006).

Friedman test with Holm's post-hoc procedure The Friedman test (Friedman 1937) is a non-parametric version of the ANOVA test. For more concrete description of how the test performs, assume we aim to compare the performance of the calibration methods in terms of $RMSE$ and our base classifier is LR. The Friedman test ranks the RMSE of LR in addition to the RMSE of the calibration methods (Hist, Platt, IsoReg, BBQ) for each dataset separately, with the best performing method getting the rank of 1, the second best the rank of 2, and so on. In case of ties, average ranks

[4] The datasets used were as follows: spect, breast, adult, page-blocks, pendigits, ad, mamography, satimage, australian, code rna, colon cancer, covtype, letter unbalanced, letter balanced, diabetes, duke, fourclass, german numer, gisette scale, heart, ijcnn1, ionosphere scale, liver disorders, mushrooms, sonar scale, splice, svmguide1, svmguide3, coil2000, balance.

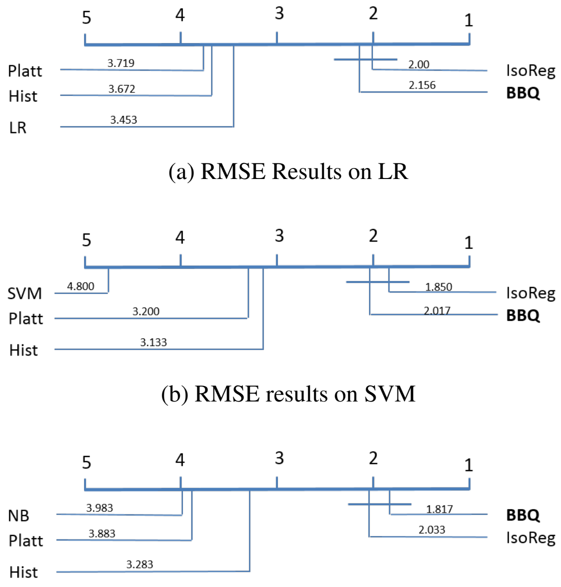


(a) ACC Results on LR

(b) ACC results on SVM

(c) ACC results on NB

Figure 4: Performance of each method in terms of average rank of ACC on the real datasets. There is no statistically significant difference between the performance of the methods in terms of ACC (using the Friedman test at a 0.05 significance level).



(a) RMSE Results on LR

(b) RMSE results on SVM

(c) RMSE results on NB

Figure 5: Performance of each method in terms of average rank of RMSE on the real datasets. All the methods which are not connected to BBQ by the horizontal bar are significantly different from BBQ (using the Friedman test followed by Holm's step-down procedure at a 0.05 significance level).

are assigned to the corresponding methods. Let $r_{i,j}$ be the rank of i'th of the 5 methods (LR, Hist, Platt, Isoreg, BBQ) at the j'th of the 30 datasets. The Friedman test computes the average rank of each method $R_i = \frac{1}{30}\sum_{j=1}^{30} r_{i,j}$. The null hypothesis states that all the methods are statistically equivalent and so their associated rank R_i should be equal. Under the null-hypothesis, the Friedman statistic

$$\chi_F^2 = \frac{12N}{k(k+1)}\Big[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\Big]$$

is distributed according to χ_F^2 with $k-1$ degrees of freedom, where N is the number of datasets (30 in our case) and k is the number of methods (5 in our case). However, it is known that the Friedman statistic is often unnecessarily conservative; thus, we use a more accurate F_F statistic (Iman and Davenport 1980) defined as follows:

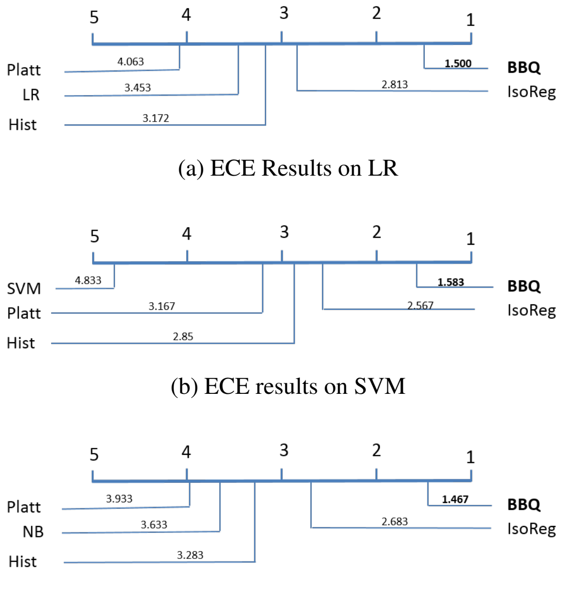
$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}$$

Under the null hypothesis the F_F statistic is distributed according to the F distribution with $k-1$ and $(k-1)(N-1)$ degrees of freedom. If the null hypothesis is rejected, we proceed with Holm's step-down post-hoc test (Holm 1979) to compare the $RMSE$ of our targeted method (BBQ in our case) to the $RMSE$ of the other methods. In order to use Holm's method, we define the z_i statistics as:

$$z_i = \frac{(R_i - R_{BBQ})}{\sqrt{\frac{k(k+1)}{6N}}},$$

where R_{BBQ} is the average rank of the target method (*BBQ*), R_i is the average rank of i'th method, k is the number of methods, and N is number of datasets. In Holm's method of testing, the z_i statistic is used to find the corresponding p_i value from the table of the normal distribution, which is compared with an adjusted α values as follows. First, the p values are sorted so that $p_{\pi_1} \leq p_{\pi_2} \ldots \leq p_{\pi_{k-1}}$. Then each p_i is compared to $\frac{\alpha}{k-i}$ sequentially. So the most significant p value, p_1, is compared with $\frac{\alpha}{k-1}$. If p_1 is below $\frac{\alpha}{k-1}$, the corresponding hypothesis is rejected and we continue to compare p_2 with $\frac{\alpha}{k-2}$, and so on. As soon as a certain null hypothesis cannot be rejected, all the remaining hypotheses are retained as well. So, if $p_j$is the first p value that is greater than $\frac{\alpha}{k-j}$, then we conclude that the rank of our target method BBQ is significantly different from the methods $\pi_1, .., \pi_{j-1}$, and it is statistically equivalent to the rest of the methods.

Results on real data sets The results on real datasets are shown in the Figures 3, 4, 5, 6, and 7. In these graphs, we indicate the average rank of each method (1 is best) and we connect the methods that are statistically equivalent with our target method *BBQ* using a horizontal bar (e.g in Figure 5a the average rank of BBQ is 2.156, it is performing statistically equivalent to IsoReg ; however, its performance in terms of RMSE is statistically superior to Hist, Platt's method, and the base classifier LR). Figure 3 shows the result of comparing the AUC of BBQ with other methods. As shown, BBQ performs significantly better than histogram binning in terms of AUC at a confidence level of

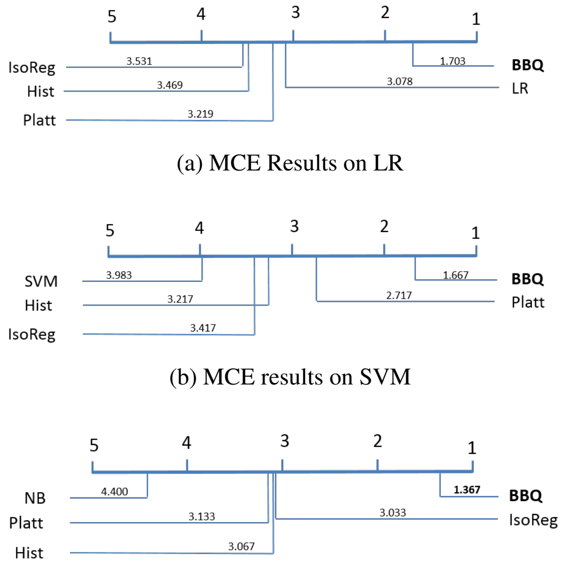


(a) ECE Results on LR

(b) ECE results on SVM

(c) ECE results on NB

Figure 6: Performance of each method in terms of average rank of ECE on the benchmark datasets. BBQ is statistically superior to all the compared methods (using the Friedman test followed by Holm's step-down procedure at a 0.05 significance level).



(a) MCE Results on LR

(b) MCE results on SVM

(c) MCE results on NB

Figure 7: Performance of each method in terms of average rank of MCE on the benchmark datasets. BBQ is statistically superior to all the compared methods (using the Friedman test followed by Holm's step-down procedure at a 0.05 significance level).

$\alpha = 0.05$. Also, its performance in terms of AUC is always statistically equivalent to the base classifier (LR, SVM, NB) and isotonic regression. Note that we did not include Platt's method in our statistical test for AUC, since the AUC of the Platt's method would be the same as the AUC of the base classifier; this pattern occurs because Platt's method always uses a monotonic mapping of the base classifier output as the calibrated scores.

Figure 4 shows the result of comparing ACC of the BBQ with the other methods. As shown, the performance of BBQ is statistically equivalent to the rest of the calibration methods as well as the base classifier in our experiments over 30 real datasets. Figure 5 shows the results of our experiments on comparing the performance of BBQ with other calibration methods in terms of RMSE. As it shows, BBQ always outperforms the base classifier, histogram binning, and Platt's method. However, its performance is statistically equivalent to isotonic regression, whether the base classifier is LR, SVM, or NB.

Figures 6 and 7 show the results of comparing BBQ performance with the others in terms of ECE and MCE, respectively. They show that BBQ performs statistically better than all other calibration methods and the base classifier, in terms of ECE and MCE.

Overall, in terms of discrimination measured by AUC and ACC, the results show that the proposed Bayesian calibration method either outperforms the other calibration methods or has a performance that is not statistically significantly different from the other methods and the base classifier. In terms of calibration performance, BBQ is statistically superior to all other methods measured by ECE and MCE. Furthermore, the results show that BBQ and isotonic regression are not statistically significantly different in terms of RMSE; however, it is still statistically superior to other calibration methods and the base classifier in terms of RMSE.

Conclusion

In this paper, we presented a Bayesian approach for Binning into Quantiles (BBQ) as a new nonparametric binary classifier calibration method. We also performed a set of experiments on simulated and real data sets to compare the discrimination and calibration performance of the method to that of other commonly applied post-processing calibration methods. The results provide support that the BBQ method performs competitively with other methods in terms of discrimination and often performs better in terms of calibration. Thus, we recommend that researchers consider using BBQ when the post-processing of binary predictions is likely to be useful.

In future work, we plan to investigate the theoretical properties of BBQ. In particular, we plan to investigate the conjecture that BBQ is expected to improve the calibration of a classifier (measured in terms of MCE and ECE) without sacrificing its discrimination capability (measured in terms of AUC). Another direction for future research is to extend BBQ to work for multi-class and multi-label calibration problems.

Acknowledgements

This research was funded in part by NIH grants R01GM088224 and R01LM010019 and NSF grant IIS-0911032.

References

Bache, K., and Lichman, M. 2013. UCI machine learning repository.

Barlow, R. E.; Bartholomew, D. J.; Bremner, J.; and Brunk, H. D. 1972. *Statistical inference under order restrictions: The theory and application of isotonic regression*. Wiley New York.

Chang, C.-C., and Lin, C.-J. 2011. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2(3):27.

DeGroot, M., and Fienberg, S. 1983. The comparison and evaluation of forecasters. *The Statistician* 12–22.

Demšar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research* 7:1–30.

Friedman, M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* 32(200):675–701.

Heckerman, D.; Geiger, D.; and Chickering, D. 1995. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning* 20(3):197–243.

Hoeting, J. A.; Madigan, D.; Raftery, A. E.; and Volinsky, C. T. 1999. Bayesian model averaging: a tutorial. *Statistical Science* 382–401.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 65–70.

Iman, R. L., and Davenport, J. M. 1980. Approximations of the critical region of the friedman statistic. *Communications in Statistics-Theory and Methods* 9(6):571–595.

Klemela, J. 2009. Multivariate histograms with data-dependent partitions. *Statistica Sinica* 19(1):159.

Menon, A.; Jiang, X.; Vembu, S.; Elkan, C.; and Ohno-Machado, L. 2012. Predicting accurate probabilities with a ranking loss. In *Proceedings of the International Conference on Machine Learning*, 703–710.

Niculescu-Mizil, A., and Caruana, R. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the International Conference on Machine Learning*, 625–632.

Platt, J. C. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers* 10(3):61–74.

Russell, S., and Norvig, P. 1995. Artificial intelligence: A modern approach. *Artificial Intelligence. Prentice-Hall, Englewood Cliffs* 25.

Scott, C., and Nowak, R. 2003. Near-minimax optimal classification with dyadic classification trees. *Advances in Neural Information Processing Systems* 16.

Zadrozny, B., and Elkan, C. 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *International Conference on Machine Learning*, 609–616.

Zadrozny, B., and Elkan, C. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 694–699.