Robust Probabilistic Calibration

Stefan Rüping[1]

Fraunhofer AIS, Schloss Birlinghoven, 53754 St. Augustin, Germany**,
stefan.rueping@ais.fraunhofer.de,
WWW home page: http://www.ais.fraunhofer.de

Abstract. Probabilistic calibration is the task of producing reliable estimates of the conditional class probability $P(class|observation)$ from the outputs of numerical classifiers. A recent comparative study [1] revealed that Isotonic Regression [2] and Platt Calibration [3] are most effective probabilistic calibration technique for a wide range of classifiers. This paper will demonstrate that these methods are sensitive to outliers in the data. An improved calibration method will be introduced that combines probabilistic calibration with methods from the field of robust statistics [4]. It will be shown that the integration of robustness concepts can significantly improve calibration performance.

1 Introduction

Given a binary classification task described by an unknown probability distribution $P(X, Y)$ on an input space X and a set of labels $Y = \{-1, 1\}$, a probabilistic classifier is a function $f_{prob} : X \to [0, 1]$ that returns an estimate of the conditional class probability, i.e.

$$f_{prob}(x) \approx P(Y = 1|x).$$

This paper deals with probabilistic classification by calibrating a numerical classifier. That is, for a numerical classification function

$$cl(x) = sign(f_{num}(x))$$

the task is to find an appropriate scaling function $\sigma : \mathcal{R} \to [0, 1]$ such that

$$\sigma(f_{num}(x)) \approx P(Y = 1|x)$$

holds. A recent comparative study [1] revealed that Platt Calibration [3] and Isotonic Regression [2] are very effective probabilistic calibration techniques for a wide range of classifiers. This paper will show that learning a probabilistic classifier is sensitive to outliers in the data and that the performance of a probabilistic classification method can be improved by taking concepts of robust statistics into account.

** The work presented in this paper was done while the author was working at LS Informatik 8, Universität Dortmund

This paper is structured as follows: Section 2 will give an introduction to performance measures for probabilistic classification and existing calibration methods. In Section 3, the new contribution of this paper will be presented, an investigation of the benefits of robustification in calibration algorithms. Section 4 gives an empirical evaluation of the calibration methods and Section 5 concludes.

2 Probabilistic Classifiers

What is a good probabilistic classifier? This is not trivial to answer, because the true conditional class probability $P(Y = 1|x)$ for an observation x is not known. One requirement is that a probabilistic classifier should be well-calibrated. That is, for each interval of probabilities $[p_1, p_2]$ the probability of drawing a positive example given the classifier predicts $f_{prob}(x) \in [p_1, p_2]$ should also be in $[p_1, p_2]$. However, calibration is not sufficient because it is easy to perfectly calibrate a classifier by assigning the default probability $P(Y = 1)$ to all examples.

A better approach is to measure the error of a probabilistic classifier on a set of examples (x_i, y_i) by a loss function, e.g. the squared loss (Brier score)

$$L_2 = \frac{1}{n} \sum_i \left(f_{prob}(x_i) - \frac{1 + y_i}{2} \right)^2$$

or the cross-entropy loss

$$L_{cre} = -\frac{1}{n} \sum_i \left(\frac{y_i + 1}{2} \log f_{prob}(x_i) + \frac{1 - y_i}{2} \log(1 - f_{prob}(x_i)) \right).$$

For these two loss functions it can be shown that a small error corresponds to a small distance of the distributions $P(Y = 1|x)$ and $f_{prob}(x)$. Both losses differ in the costs they assign to high prediction errors. In the extreme case, the cross-entropy loss is infinite when f_{num} falsely predicts a probability of 0 or 1. This does not happen with the squared loss, which is upper-bounded by 1.

2.1 Probabilistic Calibration Methods

In general, a probabilistic scaler works by letting the numerical base classifier predict the examples x_i in the training set and then fitting a scaling function that maps the predicted values to probabilities in a way that is optimal with respect to the true classes y_i. Special care must be taken to avoid over-fitting the training data. Usually, the predictions that are used as input for the calibration step are generated in an intermediate cross-validation step.

Many probabilistic calibration techniques have been proposed in the literature, for example Softmax Scaling, Binning, using the classifier's precision [5], calibration by Gaussian modeling of the decision function, Beta Scaling [6], Isotonic Regression [2], and Platt Calibration [3]. The two latter methods are the most popular of these methods and will be investigated in this paper. Both methods rely on the following assumption:

Monotonicity Assumption: The true conditional class probability $P(Y = 1|x)$ is an isotonic (monotonically increasing) function of the value of the learners decision function $f_{num}(x)$.

A recent empirical study [1] investigated the dependency between learning algorithms and their optimal calibration strategy. One of the main results of the study was that Platt Calibration works well for maximum margin methods like SVMs or Boosting, which show a similar distortion in the uncalibrated probability estimates, while it is less well suited for Naive Bayes, which shows a different type of distortion. Isotonic Regression was shown to perform consistently well for large data sets, but may suffer from overfitting on small data sets.

3 Robust Probabilistic Calibration Methods

The goal of classification algorithms is to predict the class label itself, not its probability. It follows that in order to do its job, the classifier must get a good estimation of the critical region $P(Y = 1|x) \approx \frac{1}{2}$, while it is irrelevant to get a better fit on very high and very low class probabilities. The form of the decision function far away from the class boundary may be less influenced by the data than by requirements of low complexity, sparsity, or the form of hypothesis space of the learner. For example, in Figure 1 the decision function of a radial-basis SVM is plotted. One can clearly see that near the decision boundary the function resembles a straight line, while on the outside the function is very curvy. This is an effect of the sparsity of the SVM outside the margin.

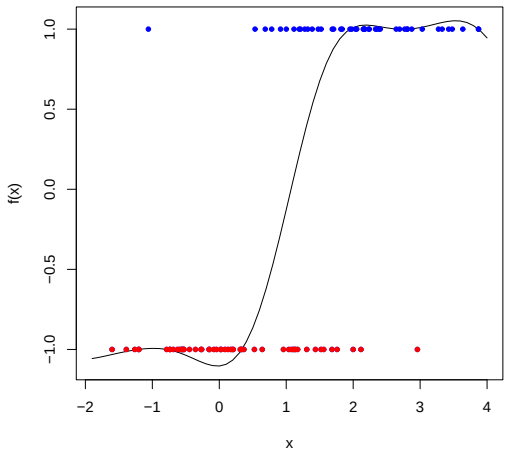

Fig. 1. Artificial one-dimensional data set and decision function of a radial-basis SVM.

The general problem of scalers which implement the monotonicity assumption in the presence of outliers is that outliers may receive extreme values of the decision function, which gives them the highest influence on the form of the scaling function. Accordingly, by removing a certain number of the points with extreme decision function values may lead to a more robust scaling function. This effect can be seen in Figure 2. It shows an one-dimensional data set, with most examples generated by two Gaussians at 0 and 2 and a small number of points generated by a Gaussian at 5. The latter points cannot be adequately modeled by the linear learner that was used and become outliers. The linear model was calibrated in two different ways, first by standard Platt Calibration and then by Platt Calibration with the outliers removed. One can clearly see that standard approach does not fit the true conditional class probability $P(Y = 1|x)$ at all, while the robustified version shows a very good fit on most of the examples.

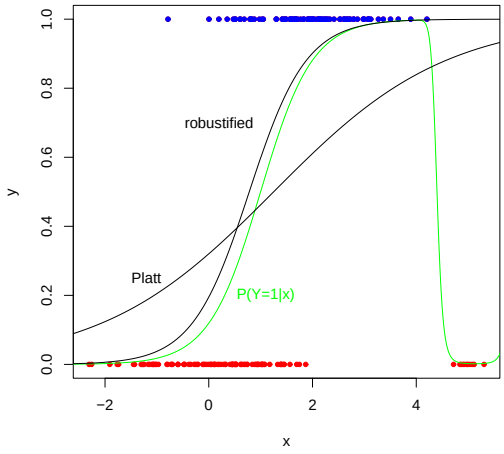

Fig. 2. Linear SVM on an artificial data set calibrated with Platt Calibration and a robustified version thereof.

This discussion raises the question whether it is advisable to drop the monotonicity assumption and allow non-monotonic scaling functions. The drawback of this idea is that this allows much more complex functions and, hence, fitting the optimal function becomes much harder. In particular, while a small set of outliers can have a drastic effect on the estimated function, it is very hard to reliably identify a better function based on only a small number of examples. However, even when there is not enough data to correctly model all examples, in many situations there at least may be enough information to remove the hazardous influence of the outlying examples.

The field of Robust Statistics [4] deals with the problem of how much influence small sets of outliers can have. Methods from this field can be used to construct a robust calibration method, i.e. one that is only minimally influenced by outliers. Even if a robust probability estimate may not be sufficient to optimally calibrate all examples, it is a practicable method to limit potential problems and to reliably identify problematic examples.

3.1 Robust Calibration

The results of the discussion so far motivate a robustification of monotonic calibration techniques by adapting the approach of least trimmed loss and least median loss [7, 8]. The main idea is to bound the influence of large outliers by removing a fraction τ of the training examples with highest absolute value $|f(x)|$, and then apply regular calibration. This approach is motivated by the least trimmed loss estimator and is based on the idea that the most deteriorating effect comes from outliers that the classifier is most sure about. Ignoring this small number of highly influential points allows the scaler to better concentrate on the bulk of the data.

In order to find an optimal value of τ, it is optimized over values in $[\tau_{min}, 1]$ for some $\tau_{min} > 0.5$. In order to get a reliable estimate of the overall error, the mean of the errors of all training examples is used to select τ. This approach, which is called Trainset-Trimmed Calibration, allows the scaler to adapt to different numbers of outliers in the data in a robust way. A lower bound of $\tau_{min} > 0.5$ is chosen in order to force the learner to predict the larger part of the data and not to fit only a small subset. In the experiments, τ_{min} was set to 0.9.

A more sophisticated, but computationally more expensive way is to select τ by means of an internal cross-validation experiment. That is, several values of τ are tried out using cross-validation of the calibration method on the training set and the optimal τ is used to calibrate to complete training set. This approach is called CV-Trimmed Calibration. Note that this internal cross-validation only requires to repeatedly execute the calibrator, not the actual learner. In the experiments, 10-fold cross-validation has been used. On the other hand, an internal cross-validation reduces the number of available examples, which may have a bad effect on small data sets.

Trimmed Calibration is independent of the underlying calibration method. In the following, Trimmed Calibration applied to Platt Calibration and Isotonic Regression will be called Trimmed Platt and Trimmed Isotonic Regression, respectively.

4 Empirical Comparison

The new probabilistic calibration algorithms have been compared with existing approaches on a total of 21 data sets, with 16 data sets from the well-known UCI machine learning repository [9], and 5 data sets from several real-world applications. The data sets were chosen as to cover a wide range of number of attributes,

Table 1. Description of the data sets used in the experiments.

Id	Name	Size	Dimension	Id	Name	Size	Dimension
1	adult	32561	104	9	iris	150	4
2	balance	576	4	10	letterP1	20000	16
3	breast-cancer	683	9	11	letterP2	20000	16
4	covtype	4951	48	12	liver	345	6
5	dermatology	184	33	13	mushroom	8124	126
6	diabetes	768	8	14	promoters	106	228
7	digits	776	64	15	voting	435	16
8	ionosphere	351	34	16	wine	178	13
17	business	157	13	20	medicine	6610	18
18	insurance	10000	135	21	garageband	1885	552
19	physics	5000	78				

dimensionality and complexity. For all data sets, continuous attributes were z-scaled, and nominal attributes have been dichotomized. Multi-class data sets have been converted to binary tasks by selecting the two largest the classes or by joining several smaller classes to one. The letter data set has been transformed into two classification tasks the same way as in [1]. For data sets with more than 5000 examples, a random sample of size 5000 has been drawn to limit the runtime of the experiments. Table 1 gives some statistics on the data sets.

The base learners in the experiments were a linear Support Vector Machine, a SVM with Radial Basis Kernel, Boosted Decision Stumps, Boosted Decision Trees, Random Forests, Decision Trees, k Nearest Neighbor and Naive Bayes. Parameters for the learners have been set to default values. All results have been obtained using 10-fold cross-validation.

In order to test the significance of the results, the Wilcoxon Signed-rank Test, which has been recently suggested in [10] for performance comparisons over multiple data sets, has been employed. In particular, it is important to compare the approaches over all data sets and all base learners combined, because it may very well be the case that in several experiments there are no outliers that distort calibration, and in this case the robustified version is intended to be identical to the standard version. Hence, finding that both methods are statistically identical in some of the cases is not considered to be a problem.

For Trainset-Trimmed Platt Calibration and both the MSE and CRE error measures the new approach is significantly better (p-values of 0.002 and 0.033) than regular Platt Calibration when compared over all bases learners (detailed results not shown here due to space constraints). However, a closer look into the results reveals that there are several learners for which it performs worse. This problem is healed by the CV-trimmed Platt Calibration, whose results are shown in Table 2; it can be seen that it not only achieves even better p-values over all learners, it also performs clearly better for most of the individual learners. The notable exception are Decision Trees and k Nearest Neighbor for the CRE measure. This is in accordance with [1], which reported that Decision

Table 2. Standard Platt Calibration versus CV-Trimmed Platt Calibration

Learner	MSE			CRE		
	% better	% worse	p-value	% better	% worse	p-value
Linear SVM	76.2	23.8	0.006	66.7	33.3	0.010
RBF SVM	76.2	23.8	0.016	57.1	28.6	0.054
Boosted Stumps	61.9	28.6	0.083	38.1	19.0	0.305
Boosted Trees	52.4	38.1	0.152	38.1	42.9	0.370
Random Forest	57.1	33.3	0.023	38.1	38.1	0.449
Decision Tree	42.9	47.6	0.848	38.1	52.4	0.866
k Nearest Neighbor	52.4	38.1	0.118	33.3	47.6	0.630
Naive Bayes	66.7	19.0	0.012	28.6	33.3	0.221
all	60.7	31.5	0.001	42.3	36.9	0.028

Trees suffer from high variance, which Platt Calibration – in contrast to Isotonic Regression – is not able to deal with. This indicates that the functional form of Platt Calibration is not adequate for Decision Trees in the first place and hence one cannot expect to get meaningful results from this combination at all. Indeed, in the following experiments we will see that robustified Isotonic Regression works clearly better for Decision Trees.

The robustification of Isotonic Regression performs slightly different than Platt Calibration. For the trainset-trimmed approach in Table 3 one can see that for both SVMs it performs exactly identical to the standard version. It is important to notice that this is not a bug, but a feature - it may very well be the case that there are no outliers which distort the calibration and hence a value of $\tau = 1$ may be optimal. In total, one can see that for every base learner the trainset-trimmed version of Isotonic Regression performs better than its standard counterpart. Over all learners, it is significantly better with a p-value of 0.002 (MSE) and 0.001 (CRE).

Surprisingly, the CV-trimmed version of Isotonic Regression performs worse than the trainset-trimmed version. An explanation can be found in the learning curve analysis in [1], which found that Isotonic Regression has much problems with overfitting when there is not enough data. Further reducing the number of available examples in an internal cross-validation scheme worsens this situation.

In summary, the mixed approach avoids a breakdown in the case of few examples, but performs worse than selecting the best robustifier for each calibrator. Best results are obtained with CV-trimmed robustification for Platt Calibration and with the Trainset-trimmed method for Isotonic Regression.

5 Conclusions

This paper presented an analysis of the robustification of the two standard probabilistic calibration techniques, Platt Calibration and Isotonic Regression. It was shown that it is possible to significantly improve both methods by safeguarding

Table 3. Standard Isotonic Regression versus Trainset-Trimmed Isotonic Regression

Learner	MSE			CRE		
	% better	% worse	p-value	% better	% worse	p-value
Linear SVM	0.0	0.0	0.500	0.0	0.0	0.500
RBF SVM	0.0	0.0	0.500	0.0	0.0	0.500
Boosted Stumps	28.6	19.0	0.238	23.8	23.8	0.270
Boosted Trees	33.3	33.3	0.450	38.1	28.6	0.265
Random Forest	52.4	9.5	0.006	52.4	9.5	0.002
Decision Tree	47.6	19.0	0.116	52.4	14.3	0.026
k Nearest Neighbor	28.6	23.8	0.175	28.6	19.0	0.093
Naive Bayes	14.3	0.0	0.091	14.3	0.0	0.091
all	25.6	13.1	0.002	26.2	11.9	0.001

against the occurrence of outliers in the data. In particular, this paper highlighted the importance of considering the concept of robustness in the design of probabilistic calibration methods.

Acknowledgments

Many thanks to Katharina Morik for her helpful advice in preparing this paper.

References

1. Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning. (2005) 625–632
2. Zadrozny, B., Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. (2002) 694–699
3. Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Smola, A., Bartlett, P., Schölkopf, B., Schuurmans, D., eds.: Advances in Large Margin Classifiers. MIT Press (1999)
4. Huber, P.J.: Robust Statistics. John Wiley & Sons (1981)
5. Rüping, S.: A simple method for estimating conditional probabilities in SVMs. In Abecker, A., Bickel, S., Brefeld, U., Drost, I., Henze, N., Herden, O., Minor, M., Scheffer, T., Stojanovic, L., Weibelza hl, S., eds.: LWA 2004 - Lernen - Wissensentdeckung - Adaptivität, Humboldt-Universität Berlin (2004)
6. Garczarek, U.: Classification Rules in Standardized Partition Spaces. PhD thesis, Universität Dortmund (2002)
7. Rousseeuw, P.J.: Least median of squares regression. J. Am. Stat. Assoc. **79** (1984) 871–880
8. Rousseeuw, P.J., Leroy, A.M.: Robust Regression and Outlier Detection. Wiley (1987)
9. Murphy, P.M., Aha, D.W.: UCI repository of machine learning databases (1994)
10. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research **7** (2006) 1–30