Multi-class probabilistic classification using inductive and cross Venn–Abers predictors

Valery Manokhin VALERY.MANOKHIN.2015@RHUL.AC.UK

Royal Holloway, University of London, Egham, Surrey, UK

Editor: Alex Gammerman, Vladimir Vovk, Zhiyuan Luo and Harris Papadopoulos

Abstract

Inductive (IVAP) and cross (CVAP) Venn–Abers predictors are computationally efficient algorithms for probabilistic prediction in binary classification problems. We present a new approach to multi-class probability estimation by turning IVAPs and CVAPs into multi-class probabilistic predictors. The proposed multi-class predictors are experimentally more accurate than both uncalibrated predictors and existing calibration methods.

Keywords: Probabilistic classification, uncertainty quantification, calibration method, on-line compression modeling, measures of confidence

1. Introduction

Multi-class classification is the problem of classifying objects into one of the more than two classes. The goal of classification is to construct a classifier which, given a new test object, will predict the class label from the set of possible k classes. In an "ordinary" classification problem, the goal is to minimize the loss function by predicting correct labels on the test set. In contrast, a probabilistic classifier outputs, for a given test object, the probability distribution over the set of k classes. Probabilistic classifiers allow to express a degree of confidence about the classification of the test object. This is useful in a range of applications and industries, including life sciences, robotics and artificial intelligence.

A number of techniques are available to solve binary classification problems. In logistic regression, the dependent variable is a categorical variable indicating one of the two classes. Logistic regression model is used to estimate the probability of binary response based on a number of independent variables (features). In large empirical comparison of supervised learning algorithms (Caruana and Niculescu-Mizil, 2006) logistic regression, whilst not competitive with the best methods, was the best model for some performance metrics in specific problems. Another technique that can be used for both binary and multi-class classification problem is Artificial neural network (ANN). Artificial neural network is a computational model based on a large collection of connected units called artificial neurons. Depending on problem setting, an ANN could be used either to predict one of the k classes directly (by having k neurons in the final layer of neural network) or indirectly by building separate networks for each of the classes. ANNs are also able estimating multi-class probabilities direct by using softmax function in the final layer of a neural-based classifier.

In a probabilistic classification setting where the loss function uses exact class probabilities, calibrating the classifier improves its performance. As shown in Caruana and Niculescu-Mizil (2005), maximum margin methods such as support vector machine and boosted trees result in sigmoid-shaped distortion of the predicted probabilities. Other methods such as neural networks and logistic regression do not suffer from these biases and result in better calibrated probabilities. The calibration method in Platt (1999) is most effective when the underlying machine learning algorithm produces sigmoid shaped distortions in the predicted probabilities, this method (called Platt's scaling) was originally developed to address distortions produced by support vector machine. Another algorithm for probability calibration is isotonic regression (Zadrozny and Elkan, 2001), this scaling method can correct any monotonic distortion. The disadvantage of isotonic regression is that it is also prone to overfitting, especially when data is scarce.

More recently, Vovk et al. (2015) introduced two new computationally efficient probabilistic predictors: IVAPs (inductive Venn–Abers predictors) and CVAPs (cross Venn–Abers predictors). IVAP can be considered a regularized form of calibration based on the isotonic regression. Due to its regularized nature, the IVAP is less prone to overfitting when compared to isotonic regression. As IVAPs are a special case of Venn–Abers predictors, they are automatically well-calibrated. CVAP is an extension of IVAP using the idea of cross-validation. In empirical studies of pairwise classification problem CVAPs have demonstrated (Vovk et al., 2015) consistent accuracy when compared to the existing methods such as isotonic regression and Platt's scaling.

All probability calibration methods described so far were designed for binary classification problems. For a multi-class classification problem there are several techniques of assigning test objects to one of the k classes. If the conditional probabilities for each of the k classes are known or can be estimated, the multi-class classification problem is reduced to a trivial task of finding the number i of the class maximizing the conditional probability $p_i(x)$ computed for each of the k classes. In practice, estimating conditional class probabilities is a hard task, especially in high-dimensional setting with limited data ("the curse of dimensionality"). In binary classification task, finding a good separating function instead of conditional probabilities often gives better prediction results.

Binary probabilistic classifiers output class probabilities or decision scores. We use Platt's scaling, IVAP and CVAP to convert output of binary classifiers into calibrated binary class probabilities. We then use the method in Section 2 to convert calibrated binary class probabilities to the multi-class probability distribution over the k classes.

The classical approach to multi-class classification is to consider a collection of binary classification problems and then combine their solutions (when solutions include pairwise class probabilities) to obtain multi-class probabilities. A number of methods for converting output of binary classifiers into multi-class probabilities are available. In a simple "one-versus-all" approach, k classifiers are built with the kth classifier separating all objects in the ith class from the objects in all other $k-1$ classes. The multi-class classifier $f(x)$ is then a function attaining $\arg\max_i f_i(x)$, where $f_i(x)$ is a classifier in binary classification problem of separating ith class from all the other classes. In another "one-versus-one" method (also called "all-pairs" classification) $\frac{k(k-1)}{2}$ binary classifiers are built to classify test objects between each pair of the ith and jth classes. If such classifiers are denoted as f_{ij}, the multi-class classification problem is reduced to finding $f_i(x)$ such that $f_i(x) = \arg\max_j \sum f_{ij}(x)$. Both "one-versus-all" and "one-versus-one" approaches generally perform well and the suitability of a method to the specific problem or application depends on the time needed to build a particular classifier in comparison with time required to repeat the classification task. For "one-versus-one" (OVO) algorithm, the number of repetitions is $O\left(N^2\right)$, whilst for "one-versus-all" (OVA) the number of repetitions is $O\left(N\right)$. However, if the time required to build a classifier is super-linear in terms of the number of objects, "one-versus-one" (OVO) is a more efficient choice.

As an alternative to solving multi-class classification by combining solutions to binary classification problems, approaches such as "single machine" and the "error correcting code" can be used. In the single machine approach (Weston and Watkins, 1999), a single optimization problem is solved by training a multi-class support vector machine to solve generalized binary support vector machine problem with the decision function $f(x) = \arg\max_{i}(w_i x + b_i)$. This approach can be used for simultaneous multi-class separation in situations where binary classification by "one-versus-all" and "one-versus-one" fails. As an additional benefit, the single machine approach results in reducing the number of support vectors and more efficient kernel computations. The benefits of the single machine are, however, limited to the situations where it is hard to separate the data whilst at the same time meaningful subsets exist which allow assigning a higher value to the decision function for the correct class as compared to other classes.

Other methods for solving multi-class classification problem include voting (Price et al., 1994) and various methods based on combining binary probabilities to obtain multi-class probabilities. Such methods rely on obtaining estimates r_{ij} of pairwise probabilities $\mu_{ij} = P(y = i \mid y \in \{i, j\}, x)$. Estimates r_{ij} are obtained by building binary classifiers for each of the pairwise unions of the ith and jth classes and using r_{ij} as an approximation for μ_{ij}. In the next section, we describe a method of converting estimates r_{ij} of pairwise class probabilities into estimates p_i of multi-class probabilities for k classes.

2. Obtaining multi-class probabilities from pairwise classification

In order to obtain estimates of multi-class probabilities we convert binary class probabilities using the method in Price et al. (1994):

$$p_i^{\mathrm{PKPD}} = \frac{1}{\sum\limits_{j:j \neq i} \frac{1}{r_{ij}} - (k-2)}. \tag{1}$$

After computing p_i^{PKPD}, the probabilities need to be normalized to ensure that they sum to one. We will refer to this method as the "PKPD" method. We use this method to obtain multi-class probabilities from pairwise class probabilities produced by applying Platt's scaling, IVAP and CVAP to the pairwise classification scores/probabilities obtained by applying underlying algorithms (we use logistics regression, support vector machine and neural network) to the test objects of each data set. We then use multi-class probabilities computed using the "PKPD" method to assign test object to one of the k classes which allows us to compute loss metrics and compare them across different calibration algorithms.

3. Inductive and cross Venn–Abers predictors

Prediction algorithms IVAP (inductive Venn–Abers predictor) and CVAP (cross Venn-Abers predictor) are computationally efficient versions of Venn-Abers predictors studied in Vovk and Petej (2014). Whilst IVAP and CVAP are based on the calibration method used by the isotonic regression (Zadrozny and Elkan, 2001), IVAP and CVAP avoid problems associated with isotonic regression such as miscalibrated probabilities or overfitting when data is scarce. As Venn–Abers predictors are a special case of Venn predictors, they inherit the property of perfect calibration from Venn predictors. As shown in Vovk et al. (2015), IVAPs are automatically perfectly calibrated and the experimental results reported in the same paper suggest that this property is inherited by CVAPs. IVAPs and CVAPs are computationally efficient algorithms with predictive efficiency depending on the efficiency of the underlying algorithms.

3.1. Computational details of IVAPs and CVAPs

IVAP uses the scores $s_1, \dots, s_k$ computed by an underlying classification algorithm on the calibration set of size k (obtained by reserving part of the training set, the other part of the training set is used for training of the underlying algorithm) and also the score s computed for a new test object. The isotonic regression is then fit twice to the set of computed scores $s_1, \dots, s_k, s$ (used as the independents variable), and two sets of dependent variable formed by combining the labels of the calibration objects with two potential labels for the test object (0 or 1 accordingly) . By fitting isotonic regression twice, IVAP computes multi-probability prediction (p_0,p_1) for the test objecte that can be interpreted as the lower and the upper probability respectively. IVAP computes (p_0,p_1) efficiently for each of the potential test objects by pre-computing two vectors F^0 and F^1 which store $f_0(s)$ and $f_1(s)$, respectively, for all possible values of s. As shown in Vovk et al. (2015), given the the scores $s_1, \dots, s_k$ of the calibration objects computed by the underlying algorithm, the IVAP's prediction rule can be computed in time $O(k \log k)$ and space $O(k)$ where k is the size of the calibration set.

A cross Venn–Abers predictor (CVAP) is just a combination of K IVAPs, where K is the number of folds in the training set. To obtain class probabilities in CVAP, Vovk et al. (2015) use minimax method to merge K multiprobability predictions by K IVAPs. For the log loss the multiprobability prediction for CVAP is an interval (1-$GM(1-p_0)$, $GM(p_1)$) obtained by computing geometric means of multiprobability predictions arising out of repeated application of IVAP to K folds ($GM(p_1)$ is the geometric mean of $p_1^1, \dots, p_1^K$ and $GM(1-p_0)$ is the geometric mean of $1-p_0^1, \dots, 1-p_0^K$). For the Brier loss, the merged probability is given by formula $p = \frac{1}{K}\sum_{k=1}^{K}\left(p_1^k + \frac{1}{2}(p_0^k)^2 - \frac{1}{2}(p_1^k)^2\right)$.

The minimax method can also be applied to IVAP to obtain single probability prediction by combining multi-probability prediction as follows: $p := p_1/(1-p_0+p_1)$.

4. Experiments on multi-class data sets

We present experimental results using several multi-class data sets: `satimage` and `vehicle silhouettes` from the Statlog collection (Michie et al., 2009), `waveform` from the UCI Machine Learning Repository (Blake and Merz, 1998) and `the mnist` (LeCun et al., 1998) . The main loss function used in the empirical studies is the log loss, defined as:

$$\text{log loss} := -y \log p. \tag{2}$$

Another popular loss function is the Brier loss.

$$\text{Brier loss} := (y - p)^2. \tag{3}$$

In both cases p is the vector of class probabilities and y is the vector of true labels one–hot encoded across the K classes. Both the log loss and the Brier loss are computed by taking arithmetic average of losses on the test set.

One advantage of the Brier loss function is that it it is still possible to compare quality of prediction in cases where prediction algorithm produces infinite log loss. In this section we compare the performance of IVAPs and CVAPs with that of Platt's scaling (Platt, 1999). We use the same underlying algorithms, namely logistic regression, neural networks and support vector machine (SVM) across all experiments. The underlying algorithms produce binary classification scores which are calibrated by applying Platt's scaling, IVAP and CVAP. The data sets and the results of the experiments are described below.

4.1. "Waveform" data set

`Waveform` is an artificial data set (Lichman, 2013) containing three different classes of waves with a total of 5,000 instances (3,500 training and 1,500 test instances) and 40 attributes. Each class is generated by combining two or three "base" waves and adding noise to each attribute. The following classification accuracies were obtained in the CART ("Classification and regression trees") study (Breiman et al., 1984): the optimal Bayes classification rate is 86% accuracy, CART decision tree algorithm — 72%, nearest neighbour algorithm — 38%.

We use the original split of the data into the training set (3,500 observations) and the test set (1,500 observations). We further split the training set into the proper training set and the validation set in proportion 3:1. To obtain pairwise classification scores, we run three underlying machine learning algorithms: support vector machine, logistic regression and neural network. We use Platt's calibration, IVAP and CVAP to convert pairwise classification scores into pairwise class probabilities. We then apply the "PKPD" (method 1) to turn pairwise classification scores into calibrated multi-class probabilities. Table 1 refers to the results of experiments.

Table 1: The Brier (top table) and log loss (bottom table) for the `waveform` data set

	Platt	IVAP	CVAP
SVM	0.3147	0.3050	0.2988
logistic regression	0.3198	0.2996	0.3003
neural network	0.3490	0.3048	0.2916
	Platt	IVAP	CVAP
SVM	0.3304	0.3075	0.3020
logistic regression	0.3316	0.2998	0.3025
neural network	0.3623	0.2973	0.2896

For all three underlying algorithms, using IVAP and CVAP to calibrate pairwise classification probabilities results in performance improvements as measured by the lower Brier and log losses. In addition, both IVAP and CVAP result in improved accuracy when compared to Platt's calibration.

4.2. "Satellite image" data set

The `Satimage` ("Landsat Satellite") data set (Lichman, 2013) contains images representing 7 different classes of soil, ranging from red or gray soil to soil containing crops such as cotton or vegetation stubble. The data set was collected to predict the soil type from new satellite images, given the multi-spectral values. The number of attributes is 36 and the number of instances is 6435.

We use the last 2,000 observations as the test set and the remaining 4,435 observations as the training set. We further split the training set into the proper training set and the validation set in proportion 3:1. Table 2 refers to the results of the experiments.

Table 2: The Brier (top table) and log loss (bottom table) for the `satimage` data set

	Platt	IVAP	CVAP
SVM	0.1538	0.1523	0.1494
logistic regression	0.2528	0.2455	0.2395
neural network	0.1710	0.1482	0.1445

	Platt	IVAP	CVAP
SVM	0.2480	0.1477	0.1445
logistic regression	0.2410	0.1491	0.1412
neural network	0.3667	0.1294	0.1265

For support vector machine and neural network, IVAP and CVAP improve on Platt's scaling in terms of both Brier and log losses. For logistic regression, IVAP and CVAP result in substantial improvement in the log loss for all three underlying algorithms as well as for Platt's scaling. For the log loss, the performance improvement by using IVAP and CVAP is quite substantial. This is due to Platt's scaling producing more "confident" probabilities for the incorrect class on some of the test points where classification errors are made. Whilst IVAP and CVAP also produce errors on some of the same test objects where Platt's scaling assigns incorrect label, IVAP and CVAP assign incorrect labels in a more cautious manner (by lowering relative probability of the incorrect class when compared to Platt's scaling). When algorithms are uncertain between two classes and Platt's scaling is assigning lower probability to the correct class (in comparison with IVAP and CVAP which assign more balanced probabilities for the uncertain cases) this results in larger penalization of Platt's scaling using log loss. This in turn is due to inherent regularization present in computing results of the application of IVAP procedure. As IVAP can be considered a regularized form of isotonic regression due to application of isotonic regression to two potential labels (0 and 1), the merged probability is averaged and is never 0 or 1.

4.3. "Vehicle silhouettes" data set

Vehicle silhouettes data set from the "Statlog collection" (Blake and Merz, 1998) was designed to find a method of distinguishing between 3D objects within a 2D image by application of an ensemble of shape feature extractors to the 2D silhouettes of the objects. Four "Corgie" model vehicles were used: "Chevrolet" van, "SAAB 9000", double–decker bus and "Open Manta 400". This particular combination of vehicles was chosen with the expectation that the bus, van and either one of the cars would be readily distinguishable, but it would be more difficult to distinguish between the cars. The data set contains 946 instances and 18 attributes.

We use 562 (2/3) observations for the training and 282 (1/3) for the testing set. we split the training set into the proper training set and the validation set in proportion 3:1. We run the same underlying algorithms and calibration methods as in all previous data sets. Table 3 refers to the experimental results for the vehicle silhouettes data set.

Table 3: The Brier (top table) and log loss (bottom table) for the vehicle silhouettes data set

	Platt	IVAP	CVAP
SVM	0.4826	0.5182	0.4563
logistic regression	0.4937	0.5099	0.4765
neural network	0.3803	0.4196	0.3423

	Platt	IVAP	CVAP
SVM	0.4773	0.5190	0.4643
logistic regression	0.4768	0.5083	0.4780
neural network	0.3728	0.4312	0.3689

Vehicle silhouettes is a complicated data set for the classification task, this is reflected in higher losses when compared to other data sets. CVAP performs better than Platt's scaling across all three underlying algorithms in terms of the Brier loss and for support vector machine and neural network in terms of the log loss.

4.4. The MNIST data set

To test the performance of multi-class probabilistic predictors based on IVAP and CVAP also on a larger data set we use the MNIST (LeCun et al., 1998) data set. The MNIST is a large database of handwritten digits commonly used for training and testing of the machine learning algorithms. The data set contains 60,000 training and 10,000 testing images.

We use the original split of the data into a training set (60,000 observations) and test set (10,000 observations). We use part of the test set (2,500 observations) for calibration. The results are shown in Table 4.

For the MNIST data set, CVAP produces Brier loss comparable to that obtained using Platt's scaling. For the log loss, both IVAP and CVAP are able to quantify log loss even in

Table 4: The Brier (top table) and log loss (bottom table) for the `MNIST` data set

	Platt	IVAP	CVAP
SVM	0.0430	0.0552	0.0498
logistic regression	0.1348	0.1355	0.1147
neural network	0.0414	0.0484	0.0435

	Platt	IVAP	CVAP
SVM	-	0.3923	0.1189
logistic regression	-	0.3430	0.2910
neural network	0.1860	0.1464	0.0949

situations where Platt's scaling results in NaNs (shown as "-" in Table 4). The formulas for combining multi-probability predictions(see Section 3.1 for more details) ensure that multi-class probabilities are never 0 or 1, this in turn guarantees that the log loss is bounded when using IVAP and CVAP.

4.5. Empirical Studies of Cross-validation

An important question is whether calibration methods perform better because of extra regularization (as one used in IVAP) or because of cross-over (when using CVAP). To investigate this, using **waveform** data set, the results for the Brier and log losses were reproduced using CVAP with different number of folds. The horizontal line is the "fold ratio" less 1, there the "fold ratio" is the ratio of the size of proper training set to the size of validation set. Figure 1 refers to the results of cross-validation experiments.

In Figure 1, support vector machine is calibrated using three different calibration algorithms, namely method of Platt (1999) (shown as "SMO" on the plot), IVAP and CVAP (Vovk et al., 2015). In terms of the Brier loss, the results for CVAP are consistently better than those for Platt's scaling. As CVAP also performs consistently better than IVAP, the results demonstrate benefits from both extra regularization and cross validation.

For logistic regression (Figure 2), the results for CVAP based on the the Brier loss are consistently better than for Platt's scaling. In terms of the log loss (Figure 2) both IVAP and CVAP deliver better performance when compared to Platt's scaling.

For neural network (Figure 3), both IVAP and CVAP perform consistently better than Platt's scaling. Similar to the cases of support vector machine (Figure 1) and logistic regression (Figure 2), when neural network is used as an underlying algorithm, CVAP performs better than IVAP.

5. Conclusion

Machine learning has made a remarkable progress. A number of methods (Caruana and Niculescu-Mizil, 2006) such as random forests, boosting, bagging and support vector machine demonstrate excellent performance exceeding the performance of earlier machine

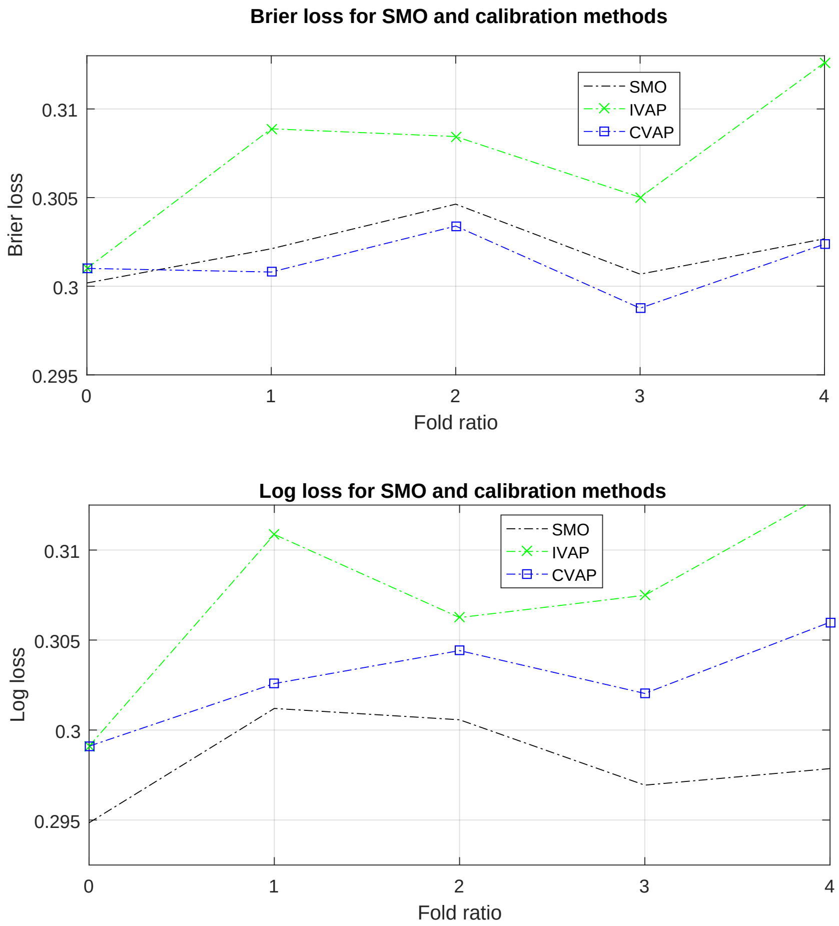


Figure 1: waveform dataset, the Brier loss (top panel) and log loss (bottom panel) for SVM calibrated with three calibrations methods.

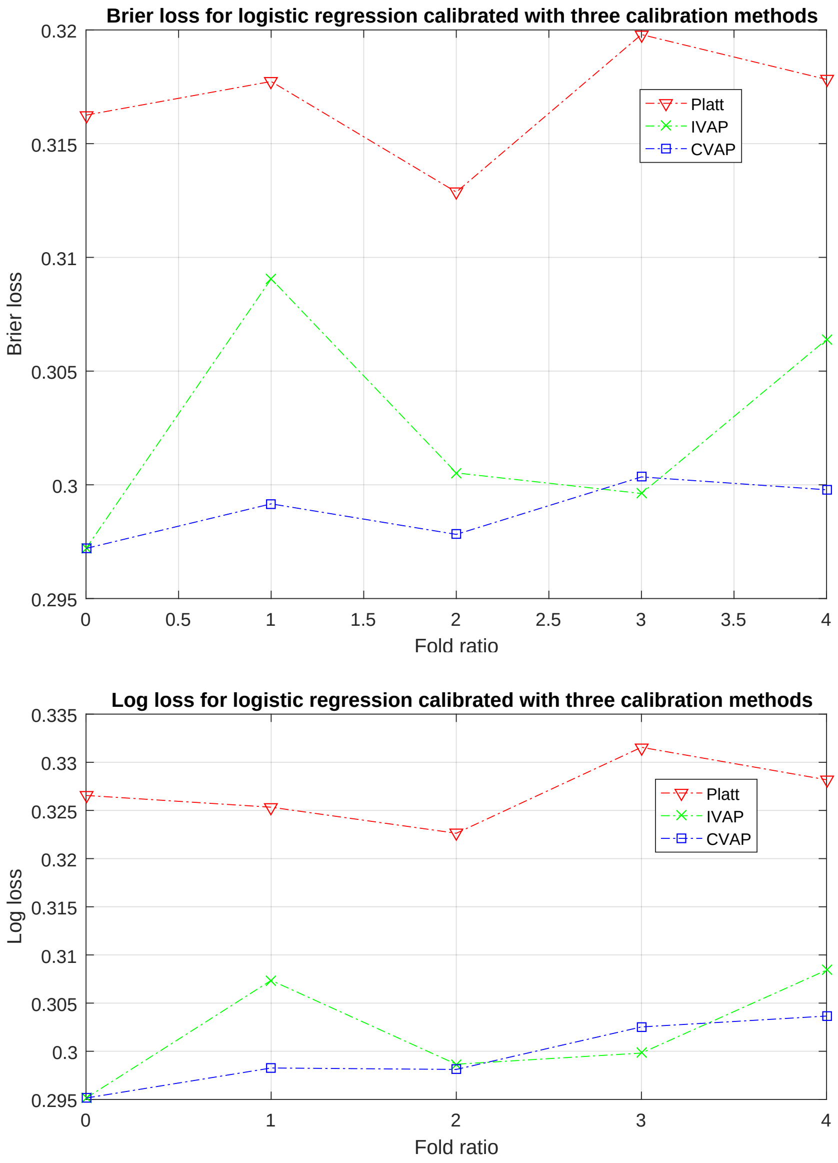
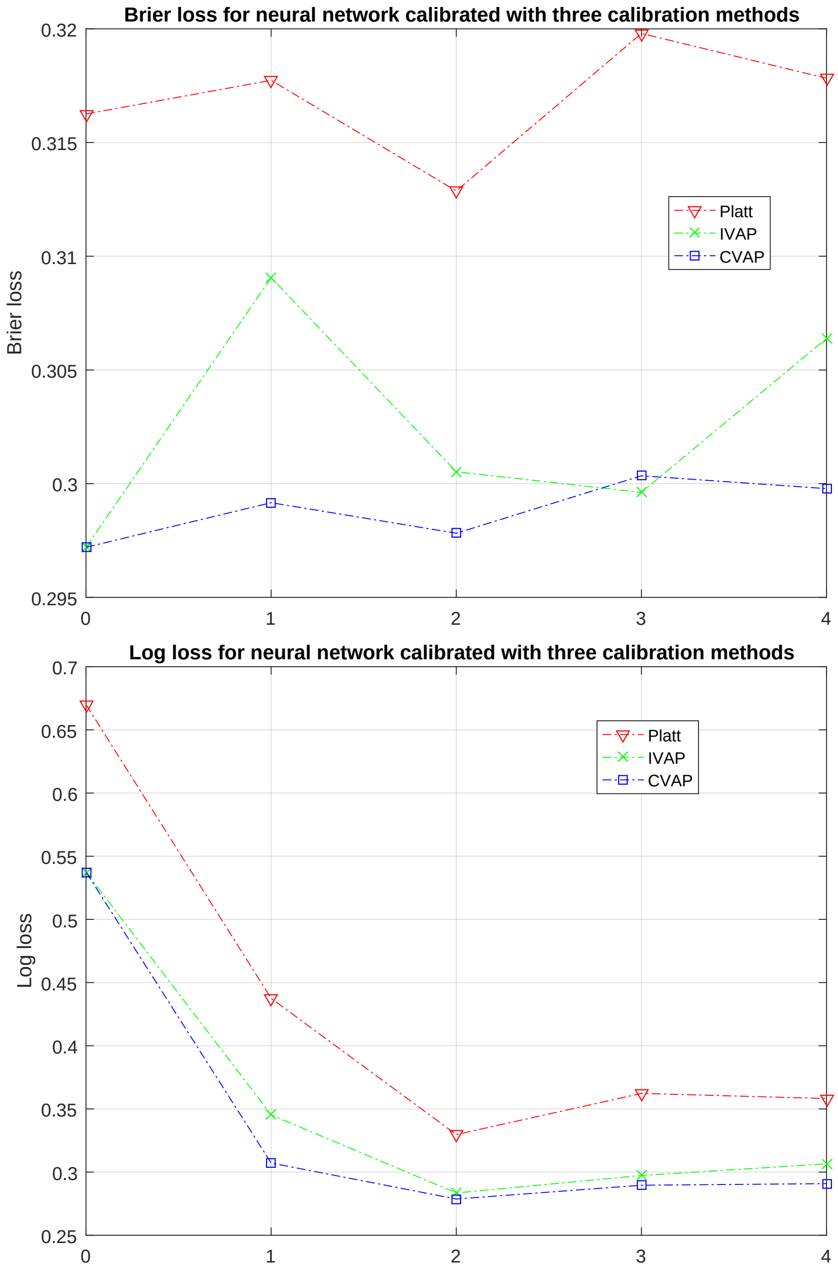


Figure 2: `waveform` dataset, the Brier loss (top panel) and log loss (bottom panel) for logistic regression calibrated with three calibrations methods.


Brier loss for neural network calibrated with three calibration methods

Log loss for neural network calibrated with three calibration methods

Figure 3: `waveform` dataset, the Brier loss (top panel) and log loss (bottom panel) for neural network calibrated with three calibrations methods.

learning algorithms. Calibration with traditional methods such as Platt's scaling or isotonic regression can improve the performance or underlying algorithms such as boosted trees, support vector machine and Naive Bayes. On the other hand, algorithms such as neural networks and logistic regression produce already well calibrated results and their performance is usually not significantly improved by calibration.

The main contribution of this paper is empirical study of the performance of two computationally efficient calibration algorithms IVAP and CVAP in the multi-class classification setting. Multi-class probability predictors based on IVAP and CVAP perform well, delivering performance improvements when compared to underlying machine-learning classification algorithms such as support vector machine, logistic regression and neural network as well as in comparison with traditional calibration method of Platt's scaling. The improvements in performance in comparison with the results produced by underlying algorithms is in line with what has been reported in the previous literature for binary classification cases, with max-margin methods such as support vector machine benefiting the most from calibration. In addition, even for well-calibrated algorithms such as neural networks and logistic regression, both IVAP and CVAP are often more accurate than the traditional calibration methods.

The additional contribution of this paper is a method of using calibration techniques for multi-class classification problems via the application of the "PKPD" method. This allows to apply well calibrated binary class probabilities in the multi-class setting in a computationally efficient manner.

Acknowledgments

Valery Manokhin is supported by the Leverhulme Magna Carta Doctoral Centre at the Royal Holloway, University of London. The author would like to thank Prof. Vladimir Vovk for his suggestion on the subject of my research and his generous and continuing help and support during my studies at Royal Holloway, University of London. The author would also like to thank anonymous reviewers for their comments and suggestions.

References

C. L. Blake and C. J. Merz. UCI repository of machine learning databases. Technical report, Department of Information and Computer Science, University of California, Irvine, CA, 1998.

Leo Breiman, Jerome Friedman, Charles J. Stone, and Richard A. Olshen. *Classification and Regression Trees*. CRC press, 1984.

Rich Caruana and Alexandru Niculescu-Mizil. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine learning*, pages 625–632. ACM, 2005.

Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 161–168, New York, NY, USA, 2006. ACM.

Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

M. Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Donald Michie, David Spiegelhalter, and Charles C. Taylor. *Machine Learning, Neural and Statistical Classification.* Overseas Press, 2009.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, pages 61–74, 1999.

David Price, Stefan Knerr, Leon Personnaz, and Gerard Dreyfus. Pairwise neural network classifiers with probabilistic outputs. In *Neural Information Processing Systems*, volume 7, pages 1109–1116, 1994.

Vladimir Vovk and Ivan Petej. Venn–Abers predictors. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, 2014.

Vladimir Vovk, Ivan Petej, and Valentina Fedorova. Large-scale probabilistic predictors with and without guarantees of validity. In *Advances in Neural Information Processing Systems*, pages 892–900, 2015.

Jason Weston and Chris Watkins. Support vector machines for multi-class pattern recognition. In *ESANN*, volume 99, pages 219–224, 1999.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *International Conference on Machine Learning*, volume 1, pages 609–616, 2001.