Applications of Class-Conditional Conformal Predictor in Multi-Class Classification

Fan Shi
Victoria Research Laboratory
National ICT Australia
Victoria, Australia
Email: fan.shi@nicta.com.au

Cheng Soon Ong
Victoria Research Laboratory
National ICT Australia
Victoria, Australia
Email: chengsoon.ong@unimelb.edu.au

Christopher Leckie
Department of Computing and Information System
The University of Melbourne
Victoria, Australia
Email: caleckie@unimelb.edu.au

***Abstract*—In many prediction problems, it is beneficial to obtain confidence estimates for the classification output. We consider the problem of estimating confidence sets in multi-class classification of real life datasets. Building on the theory of conformal predictors, we derive a class-conditional conformal predictor. This allows us to calibrate the confidence estimates in a class specific fashion, resulting in a more precise control of the prediction error rate for each class. We show that the class-conditional conformal predictor is asymptotically valid, and demonstrate that it indeed provides better calibration and efficiency on benchmark digit recognition datasets. In addition, we apply the class-conditional conformal predictor to a biological dataset for predicting localizations of proteins in order to demonstrate its performance in bioinformatics applications.**

I. Introduction

In many machine learning problems, not only the set of predictions, but also the confidence level of the predictions are needed. For example, in the application of recognising handwritten digits from 0 to 9, typical classifiers, which we refer to as *simple predictors*, output either a most likely label (i.e., 5 for a nearest-neighbor classifier) or a ranking of all labels (i.e., $(5, 7, 0, \cdots)$ for a multi-class SVM classifier) with a classifier score attached to each label for a new example. In contrast, *confidence predictors* predict a subset of ranked labels with a given confidence level, e.g., a subset of predicted labels $(5, 7)$ at a confidence level 95% usually indicates at least 95% certainty about the true label being included in the set $(5, 7)$. The confidence level can effectively inform domain experts to what extent they can trust the predictions.

An essential problem for confidence predictors is how to interpret the confidence level and its relation to the prediction error rate in a probabilistic way. A *conformal predictor*, introduced in [1], is a confidence predictor with a few additional requirements that is able to control the prediction error rate by a preset confidence level $1-\epsilon$ (or equivalently, an error rate ϵ). It essentially calibrates the output of a simple predictor into a p-value, which reflects the significance of a new example belonging to a class. Then a *prediction set* Γ^ϵ, which is a subset of class labels, can be generated according to the p-values and ϵ for the new example. The most remarkable property of the conformal predictor, which is called *validity* [1], is the ability to guarantee that the true label of the new example is included in the prediction set Γ^ϵ with a probability of at least $1-\epsilon$. In other words, when enough examples are available, the prediction error rate is no more than ϵ when we expand the set of predictions to Γ^ϵ. Validity is essential to any confidence predictors because the error rate can be effectively controlled.

However, we sometimes require the error rate in individual classes to be controlled. In the handwritten digits problem, the overall error rate may be well controlled to be no more than ϵ, but the error rate for predicting a particular digit may be significantly different. This is because the examples are drawn from distributions conditioned on the classes. This phenomenon may have severe consequences in some applications if it cannot be corrected. For example, in biological research, control of both false positives and false negatives is equally important for survival of patients in a cancer diagnostic test using genomic data.

In this paper, we propose a variant of the conformal predictor, called the *inductive class-conditional conformal predictor*, which controls the class-specific error rate without violating the overall validity. We apply the algorithm specifically to multi-class classification problem, where a significant difference in error rate between classes can be frequently observed. For this purpose, we employ the classic linear SVM [2] as the simple predictor and extend it to deal with multi-class problem using the one-against-the-rest method [3]. We also design a logistic loss-based nonconformity measure. The problem of controlling the error rate in subsets of examples was considered in [4]. This study discussed several conditional conformal predictors in general, but only verified these for the binary classification problem. In addition, the *prediction efficiency*, which is the average size of prediction set required to achieve a certain confidence level, was not evaluated.

In order to verify the performance of the class-conditional predictor, we first demonstrate the significant bias in error rate between classes for the original conformal predictor [1] on two handwritten digits datasets: MNIST [5] and USPS [6]. Then we apply the class-conditional conformal predictor to the same datasets, and illustrate the improvement in the class-specific validity and prediction efficiency. In addition, the class-conditional conformal predictor is also validated on a biological dataset for predicting localizations of proteins in bacteria, called PSORT dataset [7], [8], which illustrates its potential applications in bioinformatics studies.

In Section II, we review the conformal predictor, propose the class-conditional conformal predictor, and discuss its validity. In Section III, we present and discuss our experimental results. Finally, we conclude the paper in Section IV.

II. Methods

In this section, we first review the inductive conformal predictor proposed in [1] and its validity. Based on this, we propose an inductive class-conditional conformal predictor, which improves the control of error rates within each class. We then discuss both the overall and class-specific validity of the proposed algorithm. Finally we consider the specific choice of applying the one-against-the-rest SVMs [2], [3] as the predictor and the logistic loss as the nonconformity measure to the multi-class classification problem.

A. Inductive Conformal Predictor

Conventionally, we use $z = (x, y)$ to denote an example z comprising features x and a class label y, which belong to the feature space $\mathbb{X}$ and label space $\mathbb{Y}$, respectively. A sequence of examples with known labels $\omega := \{z_1, \cdots, z_{n-1}\}$ form the training set, while $z_n = (x_n, y_n)$ represents the test example with unknown label y_n. A general confidence predictor Γ predicts y_n with a subset of labels $\Gamma^{\epsilon}(\omega, x_n) \subseteq \mathbb{Y}$, which we call a prediction set, at a confidence level $1 - \epsilon$ (or an error rate ϵ) [9].

The inductive conformal predictor (ICP) defined in [1] splits the training set ω into a learning set $\{z_1, \cdots, z_l\}$ of size l, and a calibration set $\{z_{l+1}, \cdots, z_{n-1}\}$ of size $n-l-1$. The ICP determined by the nonconformity measure A [1] is the confidence predictor $\Gamma^{\epsilon}(\omega, x_n)$ that includes the set of all labels $y \in \mathbb{Y}$ such that

$$p_n^y > \epsilon$$

where

$$p_n^y = \frac{|\{i = l+1, \cdots, n,\ \alpha_i \geq \alpha_n\}|}{n-l} \tag{1}$$

$$\alpha_i := A(\{z_1, \cdots, z_l\}, z_i),\ i = l+1, \cdots, n-1 \tag{2}$$

$$\alpha_n := A(\{z_1, \cdots, z_l\}, (x_n, y)) \tag{3}$$

This algorithm can be interpreted from a hypothesis testing point of view. A null hypothesis is established by assuming a label $y \in Y$ for the new example: $H_0 :\ y_n = y$. The test statistic is defined as the nonconformity measure A, which measures the dissimilarity of an example (x, y) against the examples in the learning set $\{z_1, \cdots, z_l\}$. An empirical null distribution of nonconformity scores $\{\alpha_i :\ i = l+1, \cdots, n-1\}$ is formed by testing every example in the calibration set against the learning set. The score α_n for (x_n, y) is then compared with the null distribution to generate the p-value p_n^y, which is the proportion of the calibration examples that conform worse than (x_n, y). This procedure is usually called *calibration*. If the p-value p_n^y is greater than the preset threshold ϵ, then we must accept H_0. In other words, the prediction set must include y.

B. Validity of the Inductive Conformal Predictor

Intuitively, a confidence predictor Γ is valid if the probability of the true label y_n being not in the prediction set $\Gamma^{\epsilon}(\omega, x_n)$ is no more than ϵ. The formal and systematic definition of validity is given in [1]. We only review the *asymptotical validity* here. Assuming that the examples in ω are observed and predicted sequentially, the confidence predictor Γ is asymptotically valid if any exchangeable probability distribution P on Z^{∞} generating examples $\omega = (z_1, z_2, \cdots)$ and any significance level ϵ satisfy:

$$\lim_{n \to \infty} \frac{Err_n^{\epsilon}(\Gamma, \omega)}{n} \leq \epsilon \tag{4}$$

with probability one, where $Err_n^{\epsilon}(\Gamma, \omega)$ is the total number of errors made during the n examples. An error occurs if the true label of a test example is not in the prediction set. The validity of the conformal predictor was proved by formalizing Informal Proposition 1 in Appendix A in [9], which claims:

Proposition 2.1: Suppose N is large, the variables $z_1, \cdots, z_N$ are exchangeable, and E_n is an ϵ-rare event $(P(E_n|\{z_1, \cdots, z_n\}) \leq \epsilon)$ for $n = 1, \cdots, N$. Then the law of large numbers applies: with high probability, no more than approximately the fraction ϵ of the events $E_1 \cdots, E_N$ will occur.

In the ICP, the construction of the p-value (Equation 1) guarantees that making an error is an ϵ-rare event. Therefore, Proposition 2.1 can be applied to the ICP when the number of examples is large, and we immediately have the fraction of errors no more than ϵ with high probability (Equation 4).

C. Inductive Class-Conditional Conformal Predictor

Although the ICP controls the overall prediction error rate well, due to the population bias between classes, the error rate for each class could be significantly above or below the desirable ϵ even when the numbers of both training and test examples are very large. The ICP always compares the nonconformity score α_n for the new example against the same set of nonconformity measure formed by all calibration examples regardless of the tested label y. However, from the hypothesis testing point of view, when a class has its null distribution different from the combined one, testing against the common null distribution may lead to overestimates or underestimates of the p-values with respect to the class, which further affects the error rate in the class. Therefore, we introduce a conformal predictor involving a class-specific calibration procedure, which we call the *inductive class-conditional conformal predictor* (ICCCP), to counter this problem.

1) Class-specific calibration: The ICCCP is the same as the ICP except that the calibration procedure is defined as:

$$p_n^y := \frac{|\{i = l+1, \cdots, n\ :\ \alpha_i \geq \alpha_n \text{ and } y_i = y\}|}{n-l} \tag{5}$$

where α_i and α_n are the same as in (2) and (3) respectively. The major difference between the ICCCP and the ICP is that the null distribution of the nonconformity measure is class-specific in the ICCCP. For the tested label y, only the scores from the examples whose label is y are used to form the null distribution for calibration.

2) Defining the nonconformity measure A: In our algorithm, we employ the one-against-the-rest SVMs [3] as the simple predictor. Note that the linear SVM with L2-regularized and L2-loss implemented in LIBLINEAR [10] is used here due to the large scale of the benchmark datasets used in Section III, and the performance of various SVMs is not our focus. For each $y' \in \mathbb{Y}$, a binary SVM classifier discriminating class y' versus all other classes is built based on the learning set. We use y' here to avoid conflicts with the tested label y. By converting the class labels to binary values $\{-1, 1\}$ using the indicator function:

$$\Delta(y, y') = \begin{cases} 1, & \text{if } y = y' \\ -1, & \text{otherwise} \end{cases}$$

a new learning set $L^{y'}$ with respect to y' can be formed:

$$L^{y'} = \{(x_1, \Delta(y_1, y')), \cdots, (x_l, \Delta(y_l, y'))\}$$

and we use

$$D^{y'} : \mathbb{X} \to \mathbb{R}$$

to denote the binary SVM classifier trained on $L^{y'}$. Like a typical SVM classifier, it takes the features $x \in \mathbb{X}$ of an example as the input, and produces a classification score $D^{y'}(x) \in \mathbb{R}$, which is here the distance between x and the SVM hyperplane. There are $|\mathbb{Y}|$ such binary classifiers in total. A natural loss function B for these binary classifiers is the logistic loss:

$$B(L^{y'}, z_i) := \quad \frac{1}{1+\exp^{(D^{y'}(x_i)\cdot\Delta(y_i, y'))}}, \quad i = l+1, \cdots, n-1$$
$$B(L^{y'}, (x_n, y)) := \quad \frac{1}{1+\exp^{(D^{y'}(x_n)\cdot\Delta(y, y'))}}$$

When combining the measures from multiple classes in the one-against-the-rest strategy, we employ the weighted average method in [1] to define the nonconformity measure A:

$$\alpha_i := \quad \lambda B(L^{y_i}, z_i) + \tfrac{(1-\lambda)}{|\mathbb{Y}|-1} \textstyle\sum_{y'' \neq y_i} B(L^{y''}, z_i)$$
$$i = l+1, \cdots, n-1$$
$$\alpha_n := \quad \lambda B(L^{y}, (x_n, y)) + \tfrac{(1-\lambda)}{|\mathbb{Y}|-1} \textstyle\sum_{y'' \neq y} B(L^{y''}, (x_n, y))$$

where we use $\lambda = 0.5$. The choice of $\lambda \in (0, 1)$ does not have an obvious impact on our experimental observations. Using the definitions of the nonconformity measure above, we investigate some datasets in Section III.

D. Validity of the Class-Conditional Conformal Predictor

First, the above ICCCP is still asymptotically valid across all classes, which we refer to as *overall validity*. Second, the error rate in each class is also asymptotically no more than the preset threshold ϵ, which we refer to as *class-specific validity*.

Proposition 2.2: The class-conditional conformal predictor is asymptotically valid, and also asymptotically valid within each class.

In other words, Equation 4 still holds with probability one, and it is also true for the examples within each class, that is

$$\lim_{n\to\infty} \frac{Err_n^{\epsilon}(\Gamma, P, y)}{|\{i = 1, \cdots, n \ : \ y_i = y\}|} \le \epsilon, \ y \in Y$$

holds with probability one, where

$$Err_n^{\epsilon}(\Gamma, P, y) := |\{i = 1, \cdots, n \ : \ err_n^{\epsilon}(\Gamma, P) = 1 \text{ and } y_i = y\}|$$

We now discuss the overall validity of the ICCCP following the idea in Proposition 2.1. First, since the ICCCP algorithm does not make any additional assumptions on the distribution of the data, the exchangeability of the data sequence still holds. Second, although the empirical null distribution for calculating the p-value is now designed for each class (Equation 5), the decision regarding whether or not to include a label y in the prediction set is still restricted by the error rate ϵ for all examples. Therefore, the occurrence of the error $err_n^{\epsilon}(\Gamma, \omega)$ is still an ϵ-rare event. Since the ICCCP does not violate any assumptions in Proposition 2.1, its asymptotical validity across all classes still holds. Moreover, since the error rate is restricted by ϵ in each class, it is also straightforward to prove the ICCCP is asymptotically valid for each class in the same way. Thus, the class-specific error rate can be controlled using ICCCP.

III. Results

In this section, we first discuss the results of applying both the ICP and the ICCCP to two benchmark handwritten digits datasets, MNIST [5] and USPS [6] from LIBSVM [11]. The results demonstrate the practical importance and effect of correcting class bias. We then apply the two method to a biological dataset for predicting localizations of singly localized proteins in bacteria obtained from PSORTdb [7], [12], which we refer to as PSORT data, in order to illustrate the performance of the ICCCP in bioinformatics applications.

A. Evaluation Methods

When the large size and high quality of the examples are available, the theoretical properties of a conformal predictor can be well simulated by the empirical performance. Therefore, we evaluated the validity and prediction efficiency using the well-studied and large datasets USPS and MNIST. Since a fixed training set and a test set both with known labels are provided for the chosen data, we built the conformal predictor on the training set, and predict every example in the test set, and calculated the following measures.

1) Validity: Practically, when testing on a sufficiently large set, the accuracy should be close enough to the preset confidence level $1-\epsilon$ for any asymptotically valid conformal predictors. Here, the accuracy is simply defined as the proportion of test examples on which Γ^{ϵ} does not make an error. Similarly, the test accuracy can be calculated for each class separately.

2) Prediction Efficiency: Intuitively, we prefer generating prediction sets as small as possible without losing accuracy. For example, including all possible labels in a prediction set in order to gain 100% accuracy is not practically meaningful. Therefore, we measure the size of prediction set averaged on all test examples at the same confidence level. The smaller the average size of the prediction sets, the more efficient is a confidence predictor. Similarly, the prediction efficiency can also be calculated for each class separately.

B. Results on the Handwritten Digits Data

We evaluated our method on the MNIST dataset, which contains 60,000 training examples and 10,000 test examples, as well as the USPS dataset, which contains 7,291 training examples and 2,007 test examples. The ten digits from 0 to 9 represent ten class labels. We describe the results from the MNIST dataset. We repeated the same analysis for the USPS data as we did for the MNIST dataset in order to confirm our conclusions. Similar results were observed for the USPS dataset and included in the Supplement.

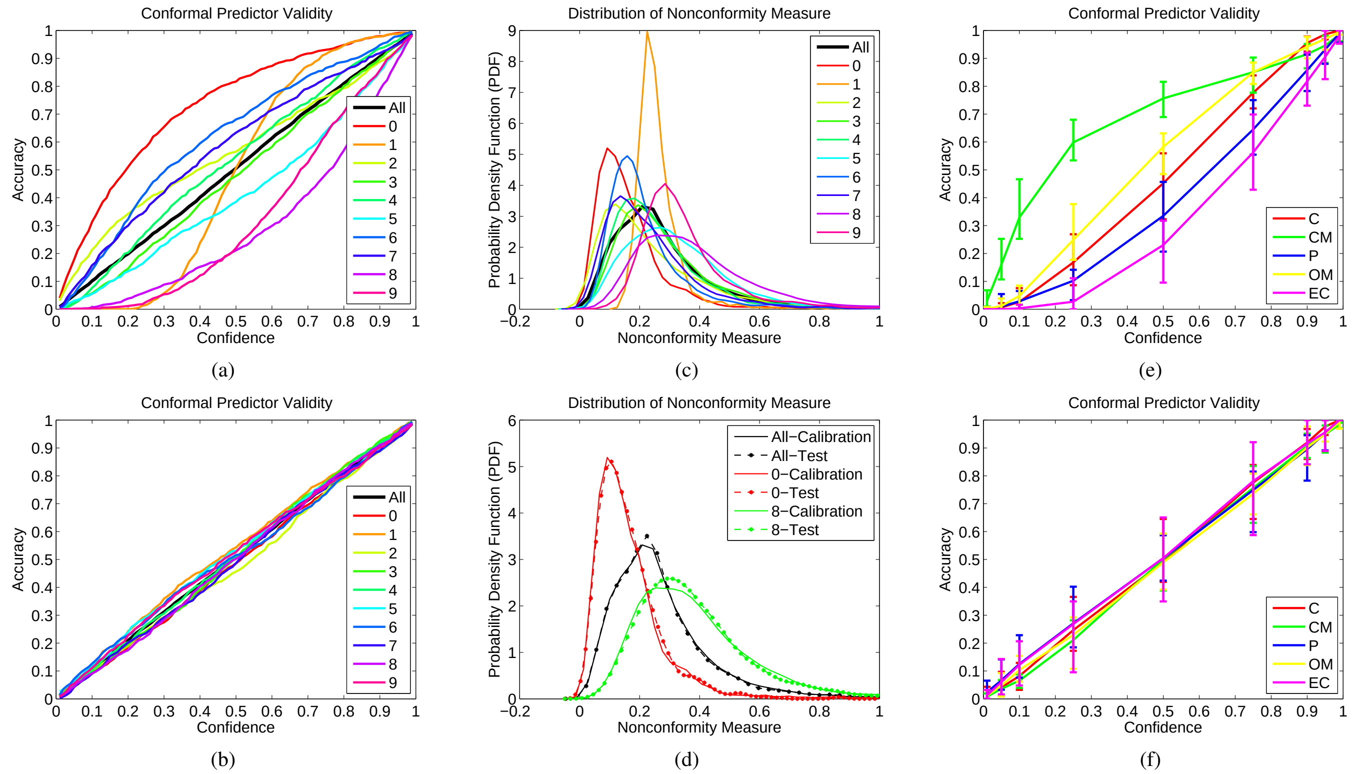

Fig. 1. (a) and (b) The overall and class-specific accuracy versus the preset confidence levels on the MNIST dataset. Each colored curve represents either the overall (black) or the class-specific confidence-accuracy relation. Figure (a) shows the results for the ICP while Figure (b) shows the results for the ICCCP. (c) The probability density function (PDF) for the distribution of the nonconformity measure on the MNIST dataset. The distribution for each class in the calibration set (colored curve) is compared with the overall distribution across all classes (black curve). Note that our nonconformity measure based on the logistic loss has no negative values. The negative values on the x-axis are caused by fitting a PDF to the histogram. (d) The fitted PDF of the nonconformity scores from both calibration and test sets are compared for the class 0, 8 and the combination of all classes on the MNIST data. (e) and (f) The class-specific accuracy versus the preset confidence levels on the PSORT dataset. The average accuracy and the error bar of the accuracy across 10 random splits are shown. Figure (e) shows the results for the ICP while Figure (f) shows the results for the ICCCP.

1) Results for the ICP: As described in Section II-A, we randomly split the training set into two halves, a learning set and a calibration set, with balanced class sizes, and built an inductive conformal predictor. We then predicted all the test examples. The settings of the classification algorithm and nonconformity measure were the same as the ICCCP in Section II-C. In addition, we have repeated the random split of the training set ten times, but the randomness did not have significant impact to the results on the benchmark datasets. Therefore, we show only the results generated in one random split for the sake of clarity. The black curve in Figure 1(a) illustrates the validity of ICP by comparing the test accuracy against a sequence of preset confidence levels $(0.01, 0.02, \cdots, 0.99)$. The closeness between the overall confidence-accuracy curve and the diagonal justifies the overall validity of ICP. However, both significant overestimates and underestimates of accuracy exist in different classes represented by colored curves.

As we discussed in Section II-C, the poor estimate of accuracy in the ICP is caused by the differences in the null distribution of the nonconformity measure between classes. This is illustrated by the probability density function (PDF) fitted to the histogram of the nonconformity scores of calibration examples in Figure 1(c). The PDFs for each class (colored curve) differ significantly from each other and also from the overall distribution (black). Thus, calibrating the test examples against the overall distribution may lead to biased estimates of the p-values and the error rate in each class.

Since we know the true labels of the test examples, we can also calculate the nonconformity measure for the test examples with respect to their true classes instead of an assumed label (note that this is not possible for real test examples with unknown labels). When comparing the distribution of the nonconformity measure between the calibration set and the test set instead of between different classes, we found that, within each class, their PDFs are very close to each other. For example, we have chosen two classes, 0 and 8, whose accuracy significantly deviates from the ideal line at opposite directions (Figure 1(a)). In Figure 1(d), the PDFs from the calibration and test sets within each of the two classes are almost identical. This phenomenon was observed for every class, but we do not show every one for the sake of clarity on the graph. Thus, the significant inter-class difference and within-class similarity imply that using the class-specific null distribution of the nonconformity measure may provide a better estimate of p-values than using the combined null distribution. This observation demonstrates the necessity of applying class-conditional calibration.

2) Results for the ICCCP: The experimental settings of the ICCCP was the same as for the ICP. As we proposed in Section II-D, the ICCCP should have both overall and class-specific validity. First, Figure 1(b) demonstrates empirically that the overall validity still holds, while the accuracy for each class is significantly closer to the preset confidence levels compared with Figure 1(a). This observation verifies the improved control over the error rate in each class.

3) Control of the prediction efficiency: Although the main purpose of the ICCCP is to control the class-specific error rate, we also observed the improvement in the control of the prediction efficiency defined in Section III-A. The results are shown in the Supplement.

C. Results on PSORT Data

Subcellular localization prediction of proteins might be helpful to developing potential diagnostic, drug and vaccine targets against bacterial [7]. PSORT [7], [12] is a dataset for predicting the localization of proteins in bacteria based on their amino acid sequence. We validated the ICCCP on the PSORT data, which contains 5 classes of localization sites, namely cytoplasm (C, 278 examples), cytoplasmic membrane (CM, 309 examples), periplasm (P, 276 examples), outer membrane (OM, 391 examples) and extracellular (EC, 190 examples). Since there is no independent test set available, we randomly split the entire dataset into a learning set, a calibration set and a test set with approximately equal size 10 times. We then applied both the ICP and ICCCP to each random split. In Figure 1(e), we show the accuracy for each class generated by the ICP. Since the PSORT dataset is not as large as the handwritten digits datasets, we show the average accuracy as well as the error bar of the accuracy across 10 splits at a selected set of confidence levels $(0.99, 0.95, 0.90, 0.75, 0.5, 0.25, 0.1, 0.05, 0.01)$. The error bars cover all the accuracy in 10 splits. Despite of random fluctuations in the accuracy, which is likely due to the relatively small number of examples, the difference in the accuracy between classes is still significant, especially between CM and EC. This was also reflected by the significant difference in recall values between classes in Table 4 in [7]. After applying the ICCCP, Figure 1(f) shows that, not only the average accuracy is close the diagonal, but also the error bars of the accuracy for each class overlap with each other. It implies that the class-specific difference is not significant any more. The analysis on the PSORT data demonstrates the potential of applying the ICCCP to bioinformatics data.

IV. Conclusion

The conformal predictor is a confidence predictor with asymptotic validity, that is, the prediction error rate is no more than a preset threshold with a large number of examples. Essentially, the validity of the conformal predictor is achieved by calibrating the nonconformity measure of new examples against the distribution of the nonconformity measure of training examples. In many applications, the class-specific validity is also desired. However, this cannot be guaranteed by the original conformal predictor since the distribution of the nonconformity measure is dependent on the class. Therefore, to control the class-specific error rate, we proposed the inductive class-conditional conformal predictor (ICCCP), where the nonconformity measure is calibrated in a class-specific fashion. Both the overall and class-specific validity hold for the ICCCP. We also designed a logistic loss-based nonconformity measure and employed one-against-the-rest SVMs to implement the ICCCP specifically for multi-class prediction problem.

The necessity and performance of the ICCCP were demonstrated on two handwritten digits datasets and the PSORT biological dataset for predicting localizations of proteins in bacteria. First, significant differences in the accuracy between classes were observed by applying the ICP to the data. This was caused by the difference in the distribution of the nonconformity measure conditioned on classes. In contrast, the high similarity in the distribution between the training examples and test examples within each class enables the class-specific calibration. Subsequently, by applying the ICCCP, the class-specific validity was significantly improved, and the difference in the prediction efficiency between classes was also reduced. According to the results, we conclude that the ICCCP with class-specific calibration is especially beneficial to correcting the class-specific error rate, especially for multi-class data with significant class biases.

Acknowledgment

The authors would like to thank National ICT Australia (NICTA) for providing funds and support, and also thank Dr. Adam Kowalczyk, who is the coordinator of the Diagnostic Genomics group in NICTA, and Dr. Jeffrey Chan in the Department of Computing and Information System, University of Melbourne the for their fruitful suggestions on this study.

References

[1] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World.* Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005.

[2] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *Intelligent Systems and Their Applications, IEEE [see also IEEE Intelligent Systems]*, vol. 13, no. 4, pp. 18–28, 1998.

[3] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *Neural Networks, IEEE Transactions on*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

[4] V. Vovk, "Conditional validity of inductive conformal predictors," http://arxiv.org/abs/1209.2673, 2012.

[5] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[6] J. Hull, "A database for handwritten text recognition research," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 16, no. 5, pp. 550–554, 1994.

[7] J. L. Gardy, M. R. Laird, F. Chen, S. Rey, C. J. Walsh, M. Ester, and F. S. L. Brinkman, "PSORTb v.2.0: Expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis," *Bioinformatics*, vol. 21, no. 5, pp. 617–623, Mar. 2005.

[8] O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne, "Predicting subcellular localization of proteins based on their n-terminal amino acid sequence." *Journal of molecular biology*, vol. 300, no. 4, pp. 1005–1016, Jul. 2000.

[9] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *J. Mach. Learn. Res.*, vol. 9, pp. 371–421, Jun. 2008. [Online]. Available: http://dl.acm.org/citation.cfm?id=1390681.1390693

[10] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008.

[11] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines (version 2.31)," 2007.

[12] C. S. Ong and A. Zien, "An automated combination of kernels for predicting protein subcellular localization." *Proceedings of the 8th Workshop on Algorithms in Bioinformatics (WABI 2008)*, pp. 186–197, 2008.