Reliable Confidence Predictions Using Conformal Prediction

Henrik Linusson[1]*, Ulf Johansson[1], Henrik Boström[2], and Tuve Löfström[1]

[1] Dept. of Information Technology, University of Borås, Borås, Sweden
{henrik.linusson, ulf.johansson, tuve.lofstrom}@hb.se
[2] Dept. of Computer and Systems Sciences, Stockholm University, Kista, Sweden
henrik.bostrom@dsv.su.se

Abstract. Conformal classifiers output confidence prediction regions, i.e., multi-valued predictions that are guaranteed to contain the true output value of each test pattern with some predefined probability. In order to fully utilize the predictions provided by a conformal classifier, it is essential that those predictions are reliable, i.e., that a user is able to assess the quality of the predictions made. Although conformal classifiers are statistically valid by default, the error probability of the prediction regions output are dependent on their size in such a way that smaller, and thus potentially more interesting, predictions are more likely to be incorrect. This paper proposes, and evaluates, a method for producing refined error probability estimates of prediction regions, that takes their size into account. The end result is a binary conformal confidence predictor that is able to provide accurate error probability estimates for those prediction regions containing only a single class label.

1 Introduction

Conformal classifiers [13] are classification models that associate each of their predictions with a measure of confidence; each prediction consists of a set of class labels, and the probability of including the true class label is bounded by a predefined level of confidence. Conformal predictors are automatically valid for any exchangeable sequence of observations, in the sense that the probability of excluding the correct class label is well-calibrated by default.

Apart from validity, the key desideratum for conformal predictors is their *efficiency*, i.e., the size of the prediction regions produced should be kept small, as they limit the number of possible outputs that need to be considered. For conformal classifiers, efficiency can be expressed as a function of the number of class labels included in the prediction regions, given a specific confidence level [12].

* This work was supported by the Swedish Foundation for Strategic Research through the project High-Performance Data Mining for Drug Effect Detection (IIS11-0053) and the Knowledge Foundation through the project Big Data Analytics by Online Ensemble Learning (20120192).

In order to make use of the confidence predictions provided by conformal classifiers, it is necessary that the prediction regions are both small and reliable. The automatic validity of conformal classifiers effectively ensures their reliability for appropriate, i.e., exchangeable, data streams, and much research has been devoted to making conformal classifiers more efficient, see e.g., [2, 4–6, 8]. However, there is a need for addressing the problem of making predictions that are *simultaneously* small and reliable. The probability of making an incorrect prediction is only valid prior to making said prediction, i.e., we know the probability of the *next* prediction being incorrect. After classifying a sequence of test patterns, however, the *a posteriori* error probability of each particular prediction is dependent on its size; this can easily be seen by noting that an empty prediction region is always incorrect, whereas a prediction region containing all possible outputs is always correct.

This paper proposes a method for utilizing posterior information, i.e., the size of prediction regions produced for a sequence of test patterns, in order to more reliably estimate the error probability of singleton predictions, i.e., predictions containing only a single class label, for binary classification problems.

2 Inductive Conformal Classification

In order to output prediction sets, conformal classifiers combine a *nonconformity function*, which ranks objects based on their apparent strangeness (compared to other observations from the same domain), together with a statistical test that can potentially reject unlikely patterns.

The nonconformity function can be any function on the form $f : X^m \times Y \to R$, but is typically based on a traditional machine learning model according to

$$f\left[h_Z, (x_i, y_i)\right] = \Delta\left[h_Z\left(x_i\right), y_i\right], \tag{1}$$

where h_Z is a predictive model trained on the problem, Z, and Δ is some function that measures the prediction errors of h_Z. For binary classification problems, a common choice of error function is

$$\Delta\left[h_Z\left(x_i\right), y_i\right] = 1 - \hat{P}_{h_Z}\left(y_i \mid x_i\right). \tag{2}$$

where $\hat{P}_{h_Z}\left(y_i \mid x_i\right)$ is a probability estimate for class y_i when the model h_Z is applied on x_i.

In order to construct an inductive conformal classifier [9,10,13], the following training procedure is used:

1. Divide the training set Z into two disjoint subsets:
   - A *proper training set* Z_t.
   - A *calibration set* Z_c, where $|Z| = q$.
2. Train a classifier h (the underlying model) on Z_t.
3. Let $\{\alpha_1, \ldots, \alpha_q\} = \{f\left(h, z_i\right), z_i \in Z_c\}$.

When a new test pattern, x_j, is obtained, its output can be predicted as follows:

1. Fix a significance level $\epsilon \in (0, 1)$.
2. For each class $\tilde{y} \in Y$:
   (a) Tentatively label x_j as $(x_j, \tilde{y})$.
   (b) Let $\alpha_j^{\tilde{y}} = f\,[h, (x_j, \tilde{y})]$.
   (c) Calculate $p_j^{\tilde{y}}$ as

$$p_j^{\tilde{y}} = \frac{\left|\left\{z_i \in Z_c : \alpha_i > \alpha_j^{\tilde{y}}\right\}\right|}{q+1} + \theta_j \frac{\left|\left\{z_i \in Z_c : \alpha_i = \alpha_j^{\tilde{y}}\right\}\right| + 1}{q+1}, \quad (3)$$

   where $\theta_j \sim U[0, 1]$.
   (d) Let $\Gamma_j^{\epsilon} = \left\{\tilde{y} \in Y : p_j^{\tilde{y}} > \epsilon\right\}$.

The resulting prediction set Γ_j^{ϵ} contains the true output y_j with probability $1 - \epsilon$. An error occurs whenever $y_j \notin \Gamma_j^{\epsilon}$, and the expected number of errors made by a conformal classifier is ϵk, where k is the number of test patterns.

3 Conformal Classifier Errors

Conformal predictors are unconditional by default, i.e., while the probability of making an error for an arbitrary test pattern is ϵ, it is possible that errors are distributed unevenly amongst different natural subgroups in the test data, e.g., test patterns with different class labels [7, 11, 13]. If the output of a test pattern is easily predicted, e.g., because it belongs to the majority class, the probability of making an erroneous prediction on that test pattern might be lower than ϵ, while the opposite might be true for difficult test patterns, e.g., those belonging to the minority class. Hence, we can express the expected number of errors made by a binary conformal classifier as

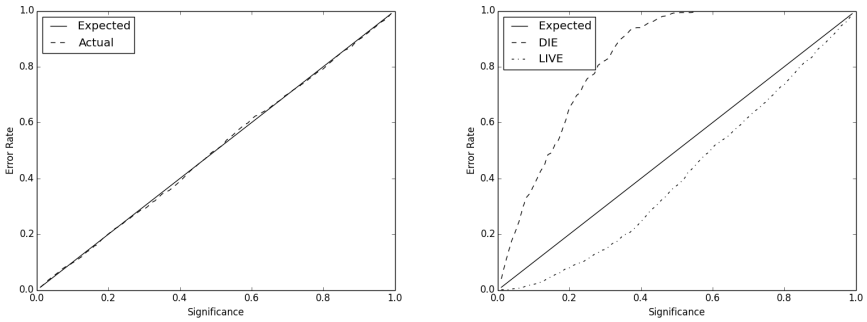
$$E = \epsilon k = \epsilon_0 k P(c_0) + \epsilon_1 k P(c_1), \quad (4)$$

where ϵ_0 and ϵ_1 are the (unknown) probabilities of making an erroneous prediction for test patterns that belong to class c_0 and c_1 respectively.

Figure 1 illustrates, using the hepatitis data set [1], the (more or less) expected behaviour of an unconditional conformal classifier for binary classification problems where the two classes are of unequal difficulty. The easier (majority) class 'LIVE' shows an error rate below ϵ, while the error rate of the more difficult (minority) class 'DIE' far exceeds ϵ.

3.1 Class-Conditional Conformal Classification

Conditional (or *Mondrian*) conformal classifiers [11, 13] effectively let us fix ϵ_0 and ϵ_1 such that $\epsilon = \epsilon_0 = \epsilon_1$ by making the p-values conditional on the class labels of the calibration examples and test patterns. This is accomplished by slightly modifying the p-value equation, so that only calibration examples that



(a) CP error rate (b) CP error rate per class

Fig. 1. Error rates of a conformal classifier on the hepatitis dataset; (a) overall error rate, i.e., over all test examples; (b) error rates for test examples belonging to the two classes, 'DIE' and 'LIVE', respectively.

share output labels with the test pattern (which is tentatively labeled as $\tilde{y}$) are considered, i.e.,

$$p_j^{\tilde{y}} = \frac{\left|\left\{z_i \in Z_\kappa : \alpha_i > \alpha_j^{\tilde{y}}\right\}\right|}{|Z_\kappa| + 1} + \theta_j \frac{\left|\left\{z_i \in Z_\kappa : \alpha_i = \alpha_j^{\tilde{y}}\right\}\right| + 1}{|Z_\kappa| + 1}, \tag{5}$$

where $Z_\kappa = \{(x_i, y_i) \in Z_c : y_i = \tilde{y}\}$ and $\theta_j \sim U[0, 1]$.

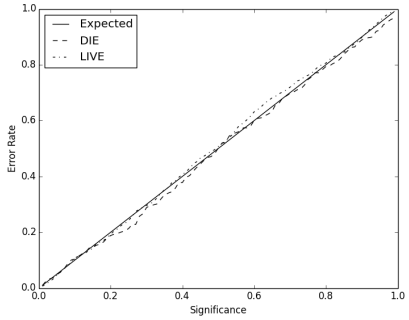


Fig. 2. Error rates of a class-conditional conformal classifier for the two classes, 'DIE' and 'LIVE', on the hepatitis data set.

Figure 2 shows the error rates of a class-conditional conformal classifier for the two classes of the hepatitis dataset. Here, a much more preferable behaviour

is observed: the error rate of the 'DIE' and 'LIVE' classes both correspond well to the expected error rate ϵ.

3.2 Utilizing Posterior Information

The overall error probability of a conformal classifier is ϵ, and class-conditional conformal classifiers extend this guarantee to apply to each class individually such that (for a binary classification problem) $\epsilon = \epsilon_0 = \epsilon_1$. This effectively handles the issue of making sure that conformal predictors can provide us with reliable predictions, regardless of class (im)balance. However, we have yet to address the task of making reliable predictions that are also small.

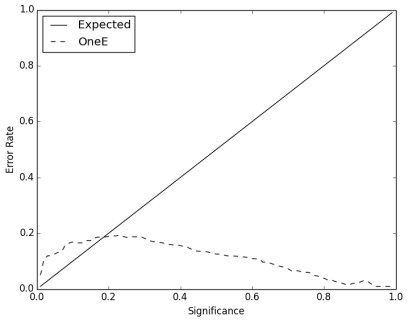


Fig. 3. OneE (error rate on singleton predictions) of a class-conditional conformal classifier on the hepatitis dataset.

For a binary classification problem, the most interesting predictions are, arguably, those containing only a single class label, i.e., the singleton predictions, since empty predictions and double predictions provide us with little actionable information. As illustrated by Figure 3, conformal classifiers, unfortunately, provide no guarantees regarding the error rate of singleton predictions; as can be seen, for the hepatitis data set, the error rate of singleton predictions (OneE) is substantially greater than ϵ for low values of ϵ.

Hence, we would like some way of expressing the likelihood of a singleton prediction being correct, without requiring knowledge of the true labels of the test patterns. To accomplish this, we are required to slightly shift our point-of-view: rather than guaranteeing the probability of making an erroneous prediction, we need to express the probability of *having made* an erroneous prediction. In the case of a binary classification problem, once k predictions have been made, we can state the expected number of errors as

$$E = \epsilon k = \epsilon k \left(P(e) + P(d) + P(s) \right), \tag{6}$$

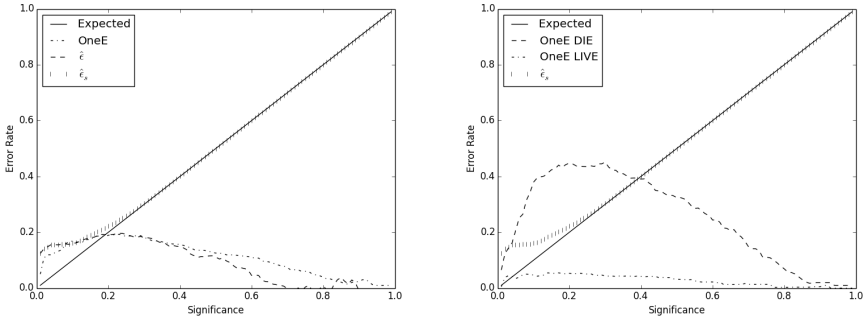
where $P(e)$, $P(d)$ and $P(s)$ are the probabilities of making empty, double and singleton predictions respectively. It is clear that we are required to make predictions (at any significance level ϵ) in order to estimate these probabilities, however, we are not required to know the true output labels of the test patterns. Once values for $P(e)$, $P(d)$ and $P(s)$ have been found, we can leverage three pieces of information regarding conformal classifiers and their prediction regions: the overall error rate on the k test patterns is ϵ; double predictions are never erroneous; and, empty predictions are always erroneous. This lets us state the following,

$$\epsilon k = \hat{\epsilon} k P(s) + k P(e) \Rightarrow \hat{\epsilon} = \frac{\epsilon - P(e)}{P(s)}, \tag{7}$$

where $\hat{\epsilon}$ is the expected error rate of the $kP(s)$ singleton predictions made. Alternatively, we can define a smoothed estimate,

$$\hat{\epsilon}_s = \frac{\epsilon}{P(s) + P(e)} \geq \sup\left\{\epsilon, \hat{\epsilon}\right\}, \tag{8}$$

where the confidence in a singleton prediction is never allowed to exceed $1 - \epsilon$.



(a) CCP OneE, with correction (b) CCP OneE per class, with correction

Fig. 4. OneE of a class-conditional conformal classifier on the hepatitis data set, with corrected ($\hat{\epsilon}$, Equation 7) and smoothed corrected ($\hat{\epsilon}_s$, Equation 8) singleton error rate estimates: (a) OneE over all test patterns; (b) OneE over test patterns belonging to the 'DIE' and 'LIVE' classes, respectively.

Figure 4 shows, again using the hepatitis data set, that the estimates $\hat{\epsilon}$ and $\hat{\epsilon}_s$ correspond well with the observed error rates on singleton predictions. From Figure 4a, it is clear that both estimates are better indicators for the OneE scores than the significance level ϵ, however, Figure 4b displays an obvious issue with both estimates: singleton predictions that indicate that the true class label is 'DIE' are incorrect much more often than expected from both $\hat{\epsilon}$ and $\hat{\epsilon}_s$, while the opposite is true for singleton predictions consisting only of the 'LIVE' class

label. Thus, it seems that we have effectively undone the efforts in making sure that the overall error rates are equal for both classes. Indeed, we would ideally want to express a reliable confidence estimate in singleton predictions for each class separately, and thus need to expand on our definition of $\hat{\epsilon}$.

For our binary classification problem, we can write the expected error rate for examples belonging to class c_i as

$$\epsilon_i = P(s_{j\neq i} \mid c_i) + P(e \mid c_i), \tag{9}$$

where, $P(s_{j\neq i} \mid c_i)$ is the probability of (erroneously) making a singleton prediction that does not include the true class c_i, and $P(e \mid c_i)$ is the probability of producing an (automatically incorrect) empty prediction for test patterns belonging to class c_i. From this we can obtain

$$\epsilon_i = P(s_{j\neq i} \mid c_i) + P(e \mid c_i) = \frac{P(c_i \mid s_{j\neq i})P(s_{j\neq i})}{P(c_i)} + P(e \mid c_i) \tag{10}$$

$$P(c_i \mid s_{j\neq i}) = \frac{P(c_i)\left[\epsilon_i - P(e \mid c_i)\right]}{P(s_{j\neq i})}, \tag{11}$$

where $P(c_i \mid s_{j\neq i}) = P(c_{i\neq j} \mid s_j)$, i.e., the probability of a prediction region containing only class c_j being erroneous. Unfortunately, this assumes that $P(e \mid c_i)$ is known—something that requires us to obtain the true class labels of our test set—however, if we assume that no empty predictions are made, we can define the estimate

$$P(e) = 0 \Rightarrow P(c_i \mid s_{j\neq i}) = \frac{\epsilon_i P(c_i)}{P(s_{j\neq i})} \geq \frac{P(c_i)\left[\epsilon_i - P(e \mid c_i)\right]}{P(s_{j\neq i})}. \tag{12}$$

Using our previous notation, we can express the estimate

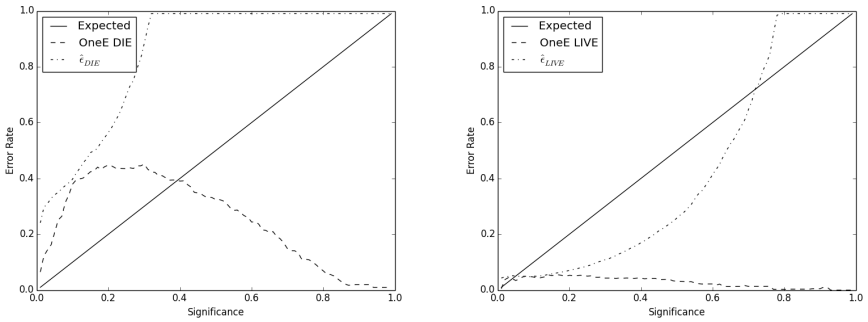
$$\hat{\epsilon}_j = \frac{\epsilon_{i\neq j} P(c_{i\neq j})}{P(s_j)}, \tag{13}$$

where $\hat{\epsilon}_j$ is the error probability of a singleton prediction containing only class c_j. It is clear that this is a conservative estimate, since the presence of empty predictions can only decrease the true expected error rate on singleton predictions. We note also that $P(c_{i\neq j})$ can be estimated from the set of calibration examples.

Figure 5, finally, displays the error rates of singleton predictions containing the 'DIE' and 'LIVE' classes, respectively, together with the estimates $\hat{\epsilon}_{DIE}$ and $\hat{\epsilon}_{LIVE}$. In both cases, the true OneE rate is approximately equal to, or lower than, the conservative estimate $\hat{\epsilon}_j$.

4 Experiments

To evaluate the proposed method of obtaining improved error rate estimates of singleton predictions, an experimental evaluation was conducted using 10x10-fold cross-validation on 20 binary classification data sets, obtained from the UCI



(a) CCP OneE DIE with correction (b) CCP OneE LIVE with correction

Fig. 5. OneE of a class-conditional conformal classifier on the hepatitis data set, with class-conditional corrected singleton error rate estimates ($\hat{\epsilon}_j$, Equation 13): (a) OneE rate for predictions containing only the 'DIE' class; (b) OneE rate for predictions containing only the 'LIVE' class.

repository [1] (Table 1). A random forest classifier [3], consisting of 300 trees, was used as the underlying model, and the calibration set size was set to 25% of the training data for all data sets. Equation 2 was used as the nonconformity function.

Table 1. Data sets used in the experiments.

Data set	#Inst.	#Feat.	#C0	#C1	Data set	#Inst.	#Feat.	#C0	#C1
balance-scale	576	5	288	288	hepatitis	155	20	32	123
breast-cancer	286	49	201	85	ionosphere	351	35	126	225
breast-w	699	10	458	241	kr-vs-kp	3196	41	1527	1669
credit-a	690	44	307	383	labor	57	27	20	37
credit-g	1000	62	300	700	liver-disorders	345	7	145	200
diabetes	768	9	500	268	mushroom	8124	122	4208	3916
haberman	306	15	225	81	sick	3772	34	3541	231
heart-c	303	23	165	138	sonar	208	61	111	97
heart-h	294	23	188	106	spambase	4601	58	2788	1813
heart-s	270	14	150	120	tic-tac-toe	958	28	332	626

Table 2 shows the rate of empty predictions (ZeroC), the rate of singleton predictions (OneC) as well as the error probability of singleton predictions (OneE) of a class-conditional conformal classifier on all 20 data sets at $\epsilon = 0.1$. Error rates in bold indicate that OneE $> 1.05\epsilon$, i.e., where the one-sided margin of error is greater than 5%. This error margin is due to the asymptotic validity of conformal predictors—we expect some statistical fluctuations in the observed

Table 2. Rate of empty predictions (ZeroC), rate of singleton predictions (OneC) and error probability of singleton predictions (OneE) of a class-conditional conformal classifier at $\epsilon = 0.1$.

$\epsilon = 0.1$	$s_0 \cup s_1$			s_0		s_1	
CCP	ZeroC	OneC	OneE	OneC	OneE	OneC	OneE
balance-scale	0.063	0.936	0.042	0.470	0.042	0.467	0.043
breast-cancer	0.000	0.346	**0.292**	0.180	**0.149**	0.166	**0.436**
breast-w	0.085	0.915	0.019	0.588	0.002	0.328	0.048
credit-a	0.004	0.912	**0.107**	0.426	**0.131**	0.486	0.085
credit-g	0.000	0.542	**0.188**	0.200	**0.355**	0.341	0.085
diabetes	0.000	0.616	**0.164**	0.378	0.089	0.238	**0.276**
haberman	0.000	0.374	**0.281**	0.238	0.105	0.136	**0.569**
heart-c	0.000	0.783	**0.127**	0.402	**0.105**	0.381	**0.144**
heart-h	0.000	0.724	**0.132**	0.424	0.064	0.300	**0.216**
heart-s	0.001	0.786	**0.130**	0.408	**0.105**	0.379	**0.149**
hepatitis	0.001	0.614	**0.169**	0.200	**0.379**	0.414	0.047
ionosphere	0.025	0.958	0.069	0.369	**0.120**	0.589	0.034
kr-vs-kp	0.098	0.902	0.001	0.431	0.001	0.471	0.002
labor	0.045	0.679	0.079	0.319	0.095	0.361	0.044
liver-disorders	0.000	0.451	**0.204**	0.239	**0.201**	0.212	**0.187**
mushroom	0.097	0.903	0.000	0.468	0.000	0.436	0.000
sick	0.087	0.913	0.014	0.845	0.001	0.068	**0.174**
sonar	0.002	0.809	**0.116**	0.446	0.097	0.363	**0.118**
spambase	0.064	0.936	0.038	0.560	0.028	0.375	0.054
tic-tac-toe	0.095	0.905	0.001	0.312	0.001	0.593	0.002
mean	0.033	0.750	0.109	0.395	0.103	0.355	0.136
min	0.000	0.346	0.000	0.180	0.000	0.068	0.000
max	0.098	0.958	0.292	0.845	0.379	0.593	0.569

error rate on a finite data set. For several of the data sets, e.g., breast-cancer, haberman and liver-disorders, the total error probability of singleton predictions ($s_0 \cup s_1$) is much greater than ϵ. This does not appear sufficient, as the singleton predictions would typically be those that are of interest to an analyst. Looking at the error rates of the individual classes, i.e., singleton predictions containing only c_0 (s_0) and singleton predictions containing only c_1 (s_1), the problem is even more pronounced—the error rate of singleton predictions containing a specific class is, for some data sets, several times greater than ϵ. So, while a conformal classifier does indeed provide us with a guarantee on the overall error probability of its predictions (when considering singleton predictions, double predictions as well as empty predictions), and even though a class-conditional conformal predictor extends this guarantee to each class separately, we cannot state any particular confidence in those prediction regions that would be of most use.

In Table 3, the same singleton error rates are tabulated, together with the the exact estimate of singleton error probability $\hat{\epsilon}$ (Equation 7), the smoothed estimate $\hat{\epsilon}_s$ (Equation 8) and the class-conditional estimate $\hat{\epsilon}_j$ (Equation 13).

Table 3. Error probabilities of singleton predictions (OneE) of a class-conditional conformal classifier at $\epsilon = 0.1$, together with estimated singleton error probabilities $\hat{\epsilon}$, $\hat{\epsilon}_s$ and $\hat{\epsilon}_j$.

$\epsilon = 0.1$	$s_0 \cup s_1$			s_0		s_1	
CCP	OneE	$\hat{\epsilon}$	$\hat{\epsilon}_s$	OneE	$\hat{\epsilon}_0$	OneE	$\hat{\epsilon}_1$
balance-scale	0.042	**0.039**	0.100	0.042	0.106	0.043	0.107
breast-cancer	0.292	0.289	0.289	0.149	0.165	0.436	0.422
breast-w	0.019	**0.017**	0.100	0.002	0.059	0.048	0.200
credit-a	0.107	0.105	0.109	0.131	0.130	0.085	0.092
credit-g	0.188	0.185	0.185	0.355	0.349	0.085	0.088
diabetes	0.164	0.162	0.162	0.089	0.092	0.276	0.273
haberman	0.281	**0.267**	0.267	0.105	0.111	0.569	0.542
heart-c	0.127	0.128	0.128	0.105	0.113	0.144	0.143
heart-h	0.132	0.138	0.138	0.064	0.085	0.216	0.213
heart-s	0.130	0.126	0.127	0.105	0.109	0.149	0.147
hepatitis	0.169	**0.161**	0.162	0.379	0.397	0.047	0.050
ionosphere	0.069	0.078	0.102	0.120	0.174	0.034	0.061
kr-vs-kp	0.001	0.002	0.100	0.001	0.121	0.002	0.101
labor	0.079	0.080	0.138	0.095	0.204	0.044	0.097
liver-disorders	0.204	0.222	0.222	0.201	0.243	0.187	0.198
mushroom	0.000	0.004	0.100	0.000	0.103	0.000	0.119
sick	0.014	0.015	0.100	0.001	0.007	0.174	1.384
sonar	0.116	0.121	0.123	0.097	0.105	0.118	0.147
spambase	0.038	0.038	0.100	0.028	0.070	0.054	0.162
tic-tac-toe	0.001	0.005	0.100	0.001	0.210	0.002	0.058

Estimates in bold indicate that OneE $> 1.05\hat{\epsilon}$. For all data sets, the exact estimate $\hat{\epsilon}$ lies close to the empirical error rate of singleton predictions. Although the estimate does exceed the true singleton error rate occasionally, we should expect it to converge with an increasing number of calibration examples and test patterns. The smoothed estimate is automatically conservative whenever the true singleton error rate is lower than the significance level ϵ, and does not substantially underestimate the true singleton error probability for any of the data sets tested on. The class-conditional estimate, $\hat{\epsilon}_j$, is often conservative, in particular for the data sets where the conformal classifier outputs a relatively large number of empty predictions, e.g., balance-scale, breast-w, kr-vs-kp; see Table 2. Again, on the data sets used for evaluation, this estimate never underestimates the singleton error probability substantially; however, for the sick data set in particular, the estimate is extremely conservative on the s_0 predictions (indicating that they are all likely to be incorrect), which is likely a result of the low rate of s_0 predictions (see Table 2).

Overall, it does indeed appear as though these three estimates are better able to more accurately express the true error probability of the singleton predictions than the original significance level ϵ. The smoothed estimate $\hat{\epsilon}_s$ and the class-conditional estimate $\hat{\epsilon}_j$, in particular, tend to overestimate rather than un-

derestimate the true singleton prediction error rate, while the exact estimate $\hat{\epsilon}$ should be expected to converge to the true error probability given enough data.

5 Concluding Remarks

In this paper, a method is proposed for providing well-calibrated error probability estimates for confidence prediction regions from a class-conditional binary conformal classifier. In particular, three estimates are proposed that express the error probability of prediction regions containing only a single class label more accurately than the original significance level, i.e., the acceptable error rate ϵ. The three estimates proposed are: an exact estimate $\hat{\epsilon}$, that expresses the error probability of singleton predictions; a smoothed estimate $\hat{\epsilon}_s$, that expresses the same probability in a conservative manner (it never falls below the original expected error rate ϵ); and, a conservative class-conditional estimate $\hat{\epsilon}_j$, that expresses the error probability of a singleton prediction containing only class c_j. All three estimates are evaluated empirically with good results.

The error probability estimates proposed in this paper do not require knowledge of the true outputs of the test set, however, it is necessary that several predictions are made before the estimates can be calibrated, as they require knowledge of the probabilities of making empty, singleton and double predictions respectively. An alternative approach, left for future work, is to obtain these probabilities from an additional validation set, or, from the calibration set itself. This could, potentially, also allow us to refine the class-conditional estimate, as it would enable us to estimate additional parameters, i.e., the probability of making an empty prediction for a test pattern belonging to a certain class, that are required to express an exact class-conditional estimate rather than a conservative one.

Another interesting direction for future work is to observe the behaviour of the proposed method in an on-line setting. As it stands, the method is best suited for use in a batch prediction setting, due to the requirement of making predictions before calculating the error probability estimates.

Finally, it would be of interest to attempt to extend the proposed method to multi-class problems as well as regression problems.

References

1. Bache, K., Lichman, M.: UCI machine learning repository. URL http://archive.ics.uci.edu/ml (2013)
2. Bhattacharyya, S.: Confidence in predictions from random tree ensembles. Knowledge and information systems 35(2), 391–410 (2013)
3. Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001)
4. Carlsson, L., Ahlberg, E., Boström, H., Johansson, U., Linusson, H.: Modifications to p-values of conformal predictors. In: Statistical Learning and Data Sciences, pp. 251–259. Springer (2015)

5. Johansson, U., Boström, H., Löfström, T.: Conformal prediction using decision trees. In: Data Mining (ICDM), 2013 IEEE 13th International Conference on. pp. 330–339. IEEE (2013)
6. Linusson, H., Johansson, U., Boström, H., Löfström, T.: Efficiency comparison of unstable transductive and inductive conformal classifiers. In: Artificial Intelligence Applications and Innovations, pp. 261–270. Springer (2014)
7. Löfström, T., Boström, H., Linusson, H., Johansson, U.: Bias reduction through conditional conformal prediction. Intelligent Data Analysis 9(6) (2015)
8. Löfström, T., Johansson, U., Boström, H.: Effective utilization of data in inductive conformal prediction using ensembles of neural networks. In: Neural Networks (IJCNN), The 2013 International Joint Conference on. pp. 1–8. IEEE (2013)
9. Papadopoulos, H.: Inductive conformal prediction: Theory and application to neural networks. Tools in artificial intelligence 18(315-330), 2 (2008)
10. Papadopoulos, H., Proedrou, K., Vovk, V., Gammerman, A.: Inductive confidence machines for regression. In: Machine Learning: ECML 2002, pp. 345–356. Springer (2002)
11. Vovk, V.: Conditional validity of inductive conformal predictors. Machine learning 92(2-3), 349–376 (2013)
12. Vovk, V., Fedorova, V., Nouretdinov, I., Gammerman, A.: Criteria of efficiency for conformal prediction. Tech. rep., Technical report, Royal Holloway University of London (April 2014) (2014)
13. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic learning in a random world. Springer Verlag, DE (2006)