Aggregated Conformal Prediction

Lars Carlsson[1], Martin Eklund[2], and Ulf Norinder[3]

[1] AstraZeneca Research and Development, SE-431 83 Mölndal, Sweden
lars.a.carlsson@astrazeneca.com
[2] Department of Surgery, University of California San Francisco (UCSF), 1600 Divisadero St, San Francisco CA 94143, USA
martin.eklund@farmbio.uu.se
[3] H. Lundbeck A/S, Ottiliavej 9, 2500 Valby, Denmark
ulfn@lundbeck.com

Abstract. We present the aggregated conformal predictor (ACP), an extension to the traditional inductive conformal prediction (ICP) where several inductive conformal predictors are applied on the same training set and their individual predictions are aggregated to form a single prediction on an example. The results from applying ACP on two pharmaceutical data sets (CDK5 and GNRHR) indicate that the ACP has advantages over traditional ICP. ACP reduces the variance of the prediction region estimates and improves efficiency. Still, it is more conservative in terms of validity than ICP, indicating that there is room for further improvement of efficiency without compromising validity.

1 Introduction

Quantitative Structure-Activity Relationship (QSAR) modeling for predicting properties, e.g. solubility or toxicity, of chemical compounds using statistical learning techniques is a widespread approach within the pharmaceutical industry to prioritize compounds for experimental testing or to alert for potential toxicity. In making informed decisions based on predictions from QSAR models, the confidence in such predictions is of vital importance and conformal predictors have been successfully applied to the drug discovery setting [1]. In particular, we have shown that a Mondrian inductive conformal predictor is efficient (i.e. informative) and almost valid when applied to binary categorical data; exact validity was not achieved due to deviations from the exchangeability assumption [2]. Furthermore, we have demonstrated that an inductive conformal predictor (ICP) applied to regression data from the pharmaceutical industry was valid. However, for regression the prediction regions were in many cases as wide or wider than the possible ranges of the true responses (i.e. the range of the experimental assay that generates the true label). An important problem is thus to improve the efficiency of the ICPs when applied in the QSAR domain.

There are many ways to improve efficiency of conformal predictions, for example by using improved nonconformity scores, choosing different machine-learning methods, or using a transductive approach [3]. The transductive approach is the

L. Iliadis et al. (Eds.): AIAI 2014 Workshops, IFIP AICT 437, pp. 231–240, 2014.

most appealing in terms of validity, but is often computationally costly and the nonconformity scores can be difficult to compute.

Another interesting approach to improve efficiency is the *cross conformal predictor* (CCP) [4]. Here the training data is divided into separate non-overlapping folds and each fold is used as a calibration set and the remainder of the data is used as a proper training set. This division allows for more data to be used for calibration and p-values are averaged over all folds. Similarly, the *bootstrap conformal predictor* (BCP) bootstraps datasets and uses the out-of-bag examples as a calibration set. p-values are then averages across all bootstrap replications.

In this paper we attempt to generalize the BCP and the CCP in what we term the *aggregated inductive conformal predictor* (ACP). We will empirically assess the ACP using data from the pharmaceutical domain and we show through a theoretical argument and experiments that ACP seems to have advantages over the standard ICP.

2 Aggregated Conformal Predictor

Consider the standard prerequisite for a description of a conformal predictor (CP), a bag of examples $\{z_1, \ldots, z_i, \ldots, z_l\}$ drawn from an exchangeable distribution Q. Each example $z_i = (x_i, y_i)$ can be described by its object $x_i \in \mathbf{X}$ and its label $y_i \in \mathbf{Y}$. The labels can be either categorical or continuous. For an inductive conformal predictor (ICP) the bag $\{z_i\}$ is partitioned into two different bags, one holding the proper training examples $\{z_1, \ldots, z_m\}$ and the other holding the calibration examples $\{z_{m+1}, \ldots, z_l\}$. The ICP p-value is then computed as

$$p = \frac{\mid \{j = m+1, ..., l : \alpha_j \geq \alpha_{l+1}\} \mid}{l - m + 1}$$

The prediction region of an ICP is determined by the "borderline" p-value, p_t, i.e. the smallest value p can obtain and still satisfy $p > \epsilon$. We can thus view p_t as the ϵth sample quantile (estimated from above)

$$p_t = U_{l-m}^{-1}(\epsilon),$$

where U_{l-m} is the empirical cumulative probability distribution of the p-values defined by

$$U_{l-m}(p) = \frac{1}{l - m + 1} \sum_{j=m+1}^{l} I(p_j < p),$$

where $I(\cdot) = 1$ if $p_j < p$ and $I(\cdot) = 0$ otherwise. We now introduce definitions of *Exchangeable resampling* and *Consistent resampling*, after which we define what we mean by an aggregated conformal predictor (ACP).

Definition 1 (Exchangeable resampling). *Let $\{z_1^*, \ldots, z_n^*\}$ be a bag of examples resampled from the empirical distribution Q_l. We call this resampling* exchangeable *if*

$$P\{(z_1^*, \ldots, z_n^*)\} = P\{(z_{\pi(1)}^*, \ldots, z_{\pi(n)}^*)\},$$

where π *is any permutation of* $\{1, ..., n\}$.

Definition 2 (Consistent resampling). *Let* $T = T(z_1, ..., z_l, Q)$ *be a statistic and* $T^* = T(z_1^*, ..., z_n^*, Q_l)$ *be an exchangeably resampled version of* T. *Further, let* G_l *and* G_l^* *be the probability distributions of* T *and* T^*, *respectively. We call the sampling process* consistent *(with respect to* T*) if*

$$\sup_z \mid G_l - G_l^* \mid \to 0 \text{ as } l \to \infty \text{ and } n \to \infty.$$

Definition 3 (ACP: Aggregated Conformal Predictor). *The following procedure is repeated* B *times, for* $b = 1, ..., B$*: Resample a bag* $\{z_1^*, \ldots, z_{n_b}^*\}$ *of examples from* $\{z_1, \ldots, z_l\}$ *using a consistent resampling procedure with respect to* α_t. *Compute the ICP p-value using the resampled bag,*

$$p_b^* = \frac{\mid \{j = m_b + 1, ..., n_b : \alpha_j^* \geq \alpha_{n_b+1}^*\} \mid}{n_b - m_b + 1}, \tag{1}$$

where α_j^* *are the nonconformity scores computed using* $\{z_1^*, \ldots, z_{n_b}^*\}$ *(*m_b *and* n_b *are indexed with* b *to make explicit that they may differ for different values of* b*). We define the ACP p-value as*

$$p_B = \frac{1}{B} \sum_{b=1}^{B} p_b^* \tag{2}$$

and the corresponding prediction region as

$$\Gamma^\epsilon(z_1, \ldots, z_l, x_{l+1}) := \{y | p_B > \epsilon\}. \tag{3}$$

A smoothed ACP can be defined analogously.

We note that the cross-conformal predictor and the bootstrapped conformal predictor suggested by Vovk [5] are two examples of ACPs.

Proposition 1. *The aggregated conformal predictor is conservatively valid.*

Proof. Since we use an exchangeable resampling procedure to construct the ACP and since an ICP is conservatively valid (Proposition 4.1, [3]), each resampled ICP in the ACP is conservatively valid (by symmetry). From this follows that the ACP also is conservatively valid.

Remark 1. Proposition 1 only holds unconditionally. The situation is different conditional on the particular dataset we have observed.

2.1 How Does the ACP Improve on the ICP?

For a p in an ICP, p_t is a hard threshold in the sense that the label y corresponding to p is either inside or outside the prediction region. Heuristically, the ACP averages over thresholds varying around p_t (since p_t^* based on a resampled bag $\{z_1^*, ..., z_l^*\}$ fluctuates around p_t), resulting in a smoothed threshold estimate with decreased variance compared to the estimate p_t.

We can use the method used by Bühlmann and Yu [6] to analyze this in a bit more detail. Consider the function

$$\delta(p) = I(p_t < p),$$

which indicates whether a p is smaller or larger than the borderline value p_t in an ICP (and thus if its corresponding label y either is inside or outside the prediction region). The sample quantile estimate p_t follows a normal distribution with mean $\epsilon = U^{-1}(\epsilon)$ and variance

$$\sigma^2 = \frac{(1-\epsilon)\epsilon}{(l-m+1)\left[f(q)\right]^2} = \frac{1-\epsilon}{(l-m+1)\epsilon},$$

where F is the population cumulative distribution function with density function f [7]. For a p in the neighborhood of ϵ

$$p = p(c) = \epsilon + c\sigma\sqrt{l-m+1} \tag{4}$$

we have the approximation

$$\delta(p(c)) \approx I(W < c), \quad W \sim N(0,1), \tag{5}$$

where W is the limiting random quantity from the asymptotic distribution of p_t (because of the construction of $p(c)$ in Equation (4)). For a fixed c, this is a hard threshold function of W. It follows that

$$\begin{aligned} &\mathrm{E}\left[\delta(p(c))\right] \to P(W < c) = \Phi(c) \text{ as } l \to \infty \text{ and } l - m \to \infty \\ &\mathrm{Var}\left[\delta(p(c))\right] \to \Phi(c)(1-\Phi(c)) \text{ as } l \to \infty \text{ and } l - m \to \infty, \end{aligned} \tag{6}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Note that the variance does not converge to zero; $\delta(p(c))$ assumes the values 0 and 1 with a positive probability even as l tends to infinity. However, for the ACP the situation looks different,

$$\begin{aligned} \delta_B(p(c)) &= \mathrm{E}^*\left[I(p_t^* \le p(c))\right] \\ &= \mathrm{E}^*\left[I\left(\sqrt{l-m+1}(p_t^* - p_t)/\sigma \le \sqrt{l-m+1}(p(c) - p_t)/\sigma\right)\right] \\ &= \Phi(\sqrt{l-m+1}(p(c) - p_t)) + o_p(1) \approx \Phi(c - W), \quad W \sim N(0,1). \end{aligned}$$

where the first approximation over the second equal sign follows because we by definition of the ACP have a consistent resampling process. Comparing with

Equation (5) for an ICP, the ACP produces a smoothed decision function of Z and therefore reduces variance. Again, following Bühlmann and Yu [6], we can study the case $p = p(0) = \epsilon$, i.e. when we are right at the population threshold and therefore has maximum variance. Then

$$\delta_B(p(0)) \to \Phi(-W) \sim U[0,1]$$

and, therefore,

$$\begin{aligned} &\mathrm{E}\left[\delta_B(p(0))\right] \to \mathrm{E}\left[U\right] = 1/2 \text{ as } l \to \infty \\ &\mathrm{Var}\left[\delta_B(p(0))\right] \to \mathrm{Var}(U) = 1/12 \text{ as } l \to \infty. \end{aligned} \tag{7}$$

Comparing Equation (6) to Equation (7), we see that the variance is reduced to one third for ACP compared to ICP.

3 Empirical Results of ACP

We used two different machine learning methods; the support vector machine (SVM) [8] implemented in the Java library version of libsvm [9] with a Gaussian radial basis kernel function

$$K(x, x') = \exp\left(-\gamma \|x - x'\|^2\right),$$

with $\gamma = 0.002$ and $C = 50$, and Random Forest (RF) [10], for which the default settings were used and each ensemble contained 100 trees.

Following Equation (16) in [11], we defined the nonconformity measure used in combination with SVM according to

$$\alpha_i = \frac{|y_i - \hat{y}_i|}{\exp\left(\hat{\mu}_i\right)}, \tag{8}$$

where $\hat{\mu}_i$ is the prediction of the value $\ln\left(|y_i - \hat{y}_i|\right)$ produced by a support vector regression machine trained on the proper training sets. After training the underlying SVM of the ICP, we calculate the residuals $|y_j - \hat{y}_j|$ for all proper training examples $j = 1, ..., m$ and train an SVM on the pairs $(x_i, \ln(|y_i - \hat{y}_i|))$. Measure (8) normalizes the absolute prediction error with the predicted accuracy of the SVM on a given example. The nonconformity measure used with RF was

$$\alpha_i = \frac{|y_i - \hat{y}_i|}{\hat{\nu}_i},$$

where $\hat{\nu}_i$ is the RF prediction of the value $|y_i - \hat{y}_i|$ with the same settings as for the other RF model.

Two public dataset (CDK5 and GNRHR) from the pharmaceutical domain was used [12]. The datasets consist of 230 and 198 examples, respectively. Each example (chemical compound) was described (characterised) by so-called signature descriptors [13] in the same way as described in [2].

The data was randomly split into two parts: A training set (80% of the original data) and a working set (20%). This procedure was repeated 50 times as to generate 50 training and working set, respectively. Furthermore, each training set was then, subsequently, randomly split into a proper training set (70% of the training set) and a calibration set (30%), similar to the 2:1 recommendation in [5]. This random selection of proper training and calibration examples was, in turn, repeated 100 times enabling the construction of 100 inductive conformal predictors for each working set. This sampling procedure is often called the $m-n$ sampling or non-replacement subsampling in the bootstrap literature and consistent with Definition 2 [14].

The results are presented in Tables 1- 8 and in Figures 1 and 2.

Table 1. The fraction of true labels within the prediction ranges for an ACP and a, for each run, randomly selected ICP at different significance levels for the 50 runs on the CDK data using SVM. Here, Q_1 and Q_3 are the lower and upper quartile, respectively.

Method	ϵ	min	Q_1	median	mean	Q_3	max
ACP	0.4	0.6522	0.7391	0.7717	0.7717	0.8043	0.8913
ICP	0.4	0.5100	0.5622	0.5997	0.5990	0.6303	0.6826
ACP	0.3	0.7174	0.8261	0.8478	0.8474	0.8696	0.9348
ICP	0.3	0.6191	0.6535	0.6879	0.6896	0.7202	0.7876
ACP	0.2	0.8043	0.8913	0.9130	0.9170	0.9565	0.9783
ICP	0.2	0.7296	0.7712	0.7923	0.7989	0.8282	0.8848
ACP	0.1	0.8913	0.9565	0.9565	0.9643	0.9783	1.0000
ICP	0.1	0.8378	0.8626	0.8933	0.8890	0.9101	0.9537

Table 2. The fraction of true labels within the prediction ranges for an ACP and a, for each run, randomly selected ICP at different significance levels for the 50 runs on the GNRHR data using SVM. Here, Q_1 and Q_3 are the lower and upper quartile, respectively.

Method	ϵ	min	Q_1	median	mean	Q_3	max
ACP	0.4	0.3590	0.5962	0.6667	0.6505	0.7179	0.8205
ICP	0.4	0.4297	0.5560	0.5960	0.5947	0.6569	0.7321
ACP	0.3	0.5128	0.6731	0.7179	0.7227	0.7628	0.8718
ICP	0.3	0.5215	0.6428	0.6736	0.6750	0.7208	0.7926
ACP	0.2	0.7179	0.7949	0.8205	0.8286	0.8462	0.9744
ICP	0.2	0.6782	0.7466	0.7749	0.7791	0.8126	0.8797
ACP	0.1	0.8205	0.8974	0.9231	0.9258	0.9487	1.0000
ICP	0.1	0.8072	0.8633	0.8745	0.8835	0.9104	0.9605

Table 3. The fraction of true labels within the prediction ranges for an ACP and a, for each run, randomly selected ICP at one significance level for the 50 runs on the CDK data using RF. Here, Q_1 and Q_3 are the lower and upper quartile, respectively.

Method	ϵ	min	Q_1	median	mean	Q_3	max
ACP	0.2	0.7174	0.8478	0.8696	0.8696	0.9130	0.9565
ICP	0.2	0.6926	0.7786	0.8071	0.7982	0.8249	0.8930

Table 4. The fraction of true labels within the prediction ranges for an ACP and a, for each run, randomly selected ICP at one significance level for the 50 runs on the GNRHR data using RF. Here, Q_1 and Q_3 are the lower and upper quartile, respectively.

Method	ϵ	min	Q_1	median	mean	Q_3	max
ACP	0.2	0.7179	0.8205	0.8718	0.8564	0.8974	1.0000
ICP	0.2	0.6854	0.7706	0.7919	0.8013	0.8271	0.9162

Table 5. The efficiency for an ACP and an ICP at one significance level for the 50 runs on the CDK5 data using SVM. Here, Q_1 and Q_3 are the lower and upper quartile, respectively.

Method	ϵ	min	Q_1	median	mean	Q_3	max
ACP	0.2	0.5874	1.2190	1.6880	3.2000	3.3220	67.6700
ICP	0.2	0.1550	0.9128	1.2620	3.5760	1.8970	398.3000

Table 6. The efficiency for an ACP and an ICP at one significance level for the 50 runs on the CDK5 data using RF. Here, Q_1 and Q_3 are the lower and upper quartile, respectively.

Method	ϵ	min	Q_1	median	mean	Q_3	max
ACP	0.2	0.4752	1.0060	1.2640	1.3750	1.5870	3.9780
ICP	0.2	0.0653	0.8659	1.1830	1.3680	1.6380	8.5950

4 Discussion

The results presented in Tables 1- 8 and in Figures 1 and 2 indicate that the ACP methodology has advantages over traditional ICP. The former adds stability and robustness to the predictions, which is particularly clear in Figures 1- 2 where the variance in the prediction ranges is smaller than from an individual ICP. This is of considerable importance within the pharmaceutical domain where precision as well as robustness in predictions are key elements for successful application in ongoing discovery projects. Although the traditional ICP is valid on average,

Table 7. The efficiency for an ACP and an ICP at one significance level for the 50 runs on the GNRHR data using SVM. Here, Q_1 and Q_3 are the lower and upper quartile, respectively.

Method	ϵ	min	Q_1	median	mean	Q_3	max
ACP	0.2	1.2770	1.625	1.733	1.907	1.903	25.830
ICP	0.2	0.5486	1.457	1.704	1.957	2.000	62.400

Table 8. The efficency for an ACP and an ICP at one significance level for the 50 runs on the GNRHR data using RF. Here, Q_1 and Q_3 are the lower and upper quartile, respectively.

Method	ϵ	min	Q_1	median	mean	Q_3	max
ACP	0.2	0.7481	1.6050	1.9620	2.1290	2.5490	6.3040
ICP	0.2	0.07138	1.3940	1.8230	2.0980	2.4840	10.3100

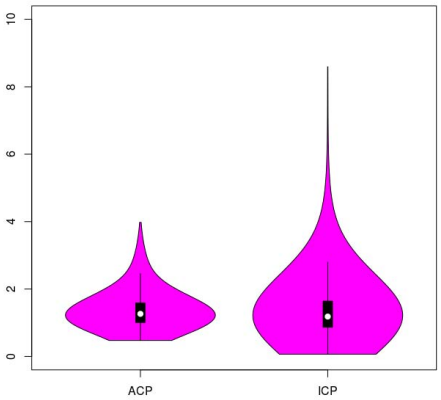

Fig. 1. The distribution of prediction interval size for an ACP and a randomly sampled ICP for all 50 runs on the CDK5 data. The results are shown at $\epsilon = 0.2$.

the variance is very large. This means that there is a relatively large proportion of very tight predictions regions that in fact are far from valid (for example, for the CDK5 data, an ICP with a confidence of 80% produces 46% error in the quartile with tightest prediction regions; the corresponding figure for the ACP is 21%. These tight prediction regions are offset by some prediction regions that are very wide (as wide or wider than the possible range of the response value) that are *always* valid, which produces a conformal predictor that is valid on average. Neither the too tight prediction regions that cannot be trusted nor non-informative regions are helpful for the researcher using the model.

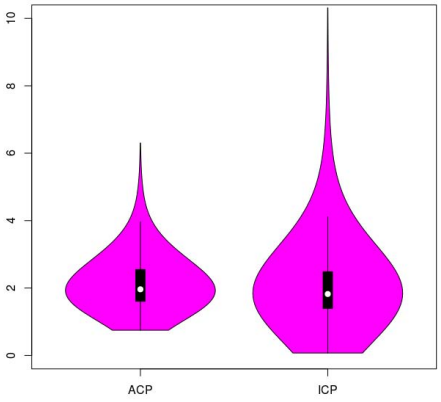

Fig. 2. The distribution of prediction interval size for an ACP and a randomly sampled ICP for all 50 runs on the GNRHR data. The results are shown at $\epsilon = 0.2$.

The results in Tables 1- 8 show that the ACP (as used in this paper) is too conservative on datasets with a relatively small number of examples, which is clear from Equations (1) and (2). This indicates that improved efficiency can be achieved without compromising the validity of the ACP, e.g. by more clever resampling (a large body of literature exists that address this problem, see e.g. [15]) or smoothing of p-values on either side of ϵ.

To conclude: We have introduced the ACP, a generalization of the BCP and CCP introduced by Vovk [5] and we have shown that it improves on classical ICP by reducing the variance in the estimated prediction regions (analogously to how a bagged predictor improves the prediction by reducing variance). ACP seems to represent a pragmatic and useful way forward for obtaining models and predictions with good precision and small prediction ranges.

Future ways to develop the ACP include to (i) study ACP for categorical labels (e.g. aggregation through voting); (ii) other resampling schemas.

References

1. Eklund, M., Norinder, U., Boyer, S., Carlsson, L.: Application of Conformal Prediction in QSAR. In: Iliadis, L., Maglogiannis, I., Papadopoulos, H., Karatzas, K., Sioutas, S. (eds.) AIAI 2012, Part II. IFIP AICT, vol. 382, pp. 166–175. Springer, Heidelberg (2012)
2. Eklund, M., Norinder, U., Boyer, S., Carlsson, L.: The application of conformal prediction to the drug discovery process. Annals of Mathematics and Artificial Intelligence (2013)
3. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World, 1st edn. Springer (2005)

4. Vovk, V.: Cross-conformal predictors. Annals of Mathematics and Artificial Intelligence, 1–20 (2013)
5. Vovk, V.: Cross-conformal predictor. Working Paper 6 (2013), http://alrw.net
6. Bühlmann, P., Yu, B.: Analyzing bagging. The Annals of Statistics 30(4), 927–961 (2002)
7. Ruppert, D.: Statistics and Data Analysis for Financial Engineering, 1st edn. Springer Texts in Statistics. Springer, Berlin (2010)
8. Vapnik, V.N.: Statistical learning theory, 1 edn. Wiley (1998)
9. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
11. Papadopoulos, H., Haralambous, H.: Reliable prediction intervals with regression neural networks. Neural Networks 24(8), 842–851 (2011)
12. Chen, H., Carlsson, L., Eriksson, M., Varkonyi, P., Norinder, U., Nilsson, I.: Beyond the scope of free-wilson analysis: Building interpretable qsar models with machine learning algorithms. Journal of Chemical Information and Modeling 53, 1324–1336 (2013)
13. Faulon, J.-L., Collins, M.J., Carr, R.D.: The signature molecular descriptor. 4. Canonizing molecules using extended valence sequences. J. Chem. Inf. Comput. Sci. 44(2), 427–436 (2004)
14. Politis, D.N., Romano, J.P., Wolf, M.: Subsampling. Springer, New York (1999)
15. Egloff, D., Leippold, M.: Quantile estimation with adaptive importance sampling. The Annals of Statistics 38(2), 1244–1278 (2010)