Proceedings of Machine Learning Research 128:1–15, 2020 Conformal and Probabilistic Prediction and Applications

Batch mode active learning for mitotic phenotypes using conformal prediction

Adam M Corrigan
Philip Hopcroft
Ana J Narvaez
Claus Bendtsen CLAUS.BENDTSEN@ASTRAZENECA.COM
Discovery Sciences, R&D, AstraZeneca, Cambridge, UK

Editor: Alexander Gammerman, Vladimir Vovk, Zhiyuan Luo, Evgueni Smirnov and Giovanni Cherubin

Abstract

Machine learning models are now ubiquitous in all areas of data analysis. As the amount of data generated continues to increase exponentially, the task of annotating sufficient objects with known labels by an expert remains expensive. To mitigate this, active learning approaches attempt to identify those objects whose labels will be most informative.

Here, we introduce a batch-based active learning framework in a pooled setting based around conformal predictors. We select objects to add to the labelled observations based on perceived novelty, while mitigating the risks of selecting highly correlated or outlying observations. We compare our approach to classical methods using an example UCI dataset, and demonstrate its application to a pharmaceutically relevant cellular imaging problem for classifying mitotic phenotypes. Our approach facilitates efficient discovery of rare and novel classes within large screening datasets.

Keywords: Active Learning, Conformal Prediction, DNA Damage Response

1. Introduction

The ongoing revolution in digital technology has meant that the generation of data is relatively cheap, even at large scales. The most expensive step when building a classification model is often the labelling of a subset of the data for training. Depending on the application, there are many ways that the labels are constructed. Often this step entails an expert examining and manually annotating the data, although the generation of labelled training data could also require performing an experiment. Naïvely, the observations selected for training will be sampled randomly from the whole dataset, or are selected by the annotator in an ad hoc manner, which can lead to redundancy within the training data. Given the cost to produce labelled data, significant resource savings can be achieved by identifying those objects which add the most value to a model and presenting only these for annotation, in an approach commonly termed active learning.

The active learning strategy adopted depends on the setting in which the data is collected. In an online or stream-based setting, where data points arrive sequentially, the active learner must decide, for each point, whether or not to present this for labelling (Settles, 2009), to make best use of the labelling resource. Alternatively, in pool-based approaches the full unlabelled dataset is available at the outset, and the task is to prioritise the datapoints for labelling often with the aim of achieving adequate model performance within a preset budget of labelling. In either setting, once an instance has been selected, its label is *opened* for annotation by the expert and the model is retrained with the updated training set.

An important choice in active learning is the strategy by which instances are selected for annotation. A common approach, referred to as *uncertainty sampling*, chooses instances which are classified with the least certainty, using the posterior probability under the current model. The degree of uncertainty can be quantified directly as the least confident predicted label, or for multi-class problems can take into account information from the probability distribution of predicted labels, by calculating the margin between highest and second-highest label probabilities (margin sampling) (Scheffer et al., 2001) or quantifying the entropy of the label distribution for each instance (Settles, 2009; Shannon, 1948). An alternative approach for instance selection calculates how the model would be changed by adding the instance, either in terms of the expected update to model parameters (Settles et al., 2008) or the expected improvement in model performance (Roy and McCallum, 2001).

In some scenarios, it is preferred that batches of unlabelled data are selected for labelling, for example when retraining of the model is costly, or where labelling can be done more efficiently in batches. A key challenge in the batch setting is the potential for redundancy across the instances in a batch if the instances are selected independently. In other words, the selection of instances within a batch does not take account of the effect of other members of the batch. To overcome this, approaches have been developed which consider the batch as a whole and choose observations that reduce overall model uncertainty by applying the Fisher information matrix (Hoi et al., 2006), maximizing the discriminative classification performance whilst also considering the overall pool (Guo and Schuurmans, 2008), maximising joint entropy (Nguyen et al., 2012), or using meta-learning (Ravi and Larochelle, 2018).

When using the performance of the model as a metric for actively selecting instances for training or determining when adequate performance has been achieved and no further instances are needed, an important consideration is the confidence with which classifications can be made. It is often the case with standard machine learning methods that the estimated posterior probabilities are not well calibrated, which implies that it is infeasible to determine an appropriate threshold a priori for candidate selection or deciding when to stop the active learning process. Conformal prediction (Vovk et al., 2014) is an approach which uses the concept of non-conformity to form so-called valid predictions, under the assumption of exchangeability. In the context of active learning, conformal prediction has previously been applied to identify new queries as those instances for which the current model has low credibility and confidence (Balasubramanian et al., 2014), for active class selection (Nouretdinov, 2017), and in the context of convolutional neural networks (Matiz and Barner, 2019).

In this work, we apply the concept of conformal prediction to pool-based active learning in batch mode. We design a scheme which selects instances to query which have low credibility whilst preserving validity of the predictions.

2. Methods

2.1. Batch Mode Active Learning Framework

Let $\mathbf{X}$ and $\mathbf{Y}$ be fixed non-empty measurable spaces; we will call them the *object* and *label spaces*, respectively. The Cartesian product $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$ is the *observation space*. Suppose we are given an initial *training* set[1] $T_0 = \lbag z_1, \ldots, z_n \rbag$ of observations, $z_i = (x_i, y_i) \in \mathbf{Z}$ and a set of objects, $S = \lbag x_{n+1}, \ldots, x_N \rbag$ with $x_i \in \mathbf{X}$. We are interested in how to select the most informative objects by sampling from the pool S for labelling and adding to the training set T_i, $i = 0, 1, \ldots$.

One classical approach for doing so, which often works well in practice, is uncertainty sampling, including the *least confident* approach (Culotta and McCallum, 2005) which proceeds iteratively, $i = 0, 1, \ldots$, based on a model $\theta_i : \mathbf{X} \rightarrow \mathbf{Y}$ trained on T_i and selects x_i^{LC} based on evaluating the posterior probability P_{θ_i} under the model such that:

$$x_i^{LC} = \text{argmin}_{x \in S} P_{\theta_i}(\text{argmax}_{y \in \mathbf{Y}} P_{\theta_i}(y)),$$

I.e. select the object from the pool for which we have the largest uncertainty in its predicted label. The label y_i^{LC} for x_i^{LC} is then opened and the training set is augmented with the new observation, $T_{i+1} = T_i \uplus \lbag z_i^{LC} \rbag$, $z_i^{LC} = (x_i^{LC}, y_i^{LC})$ and the process is repeated (margin sampling and using entropy as uncertainty measure are slight variations to this). A variation which forms the basis of the query by transduction (QbT) method is to select objects where the difference between the two largest probabilities is small (Balasubramanian et al., 2014).

From a practical perspective we are particularly interested in the case when opening of the label for objects from S are best done in batches, i.e. we only open labels when we have a set of m objects to open. The most straightforward extension to the least confident approach above for a batch setting is simply only to update the model θ_i with new training data at every m^{th} iteration, which corresponds to selecting the m least confident objects from the pool. As described in the introduction a drawback with this approach is that it will disregard any correlation between objects within a given batch and will be sensitive to outliers.

To address these drawbacks we here propose to simply replace the selection of the most uncertain objects with a random sampling from candidate objects in which we have high uncertainty overall. The purpose with this is that we refrain from prioritising amongst specific objects provided that their predicted label has low credibility, i.e. when they appear novel overall. We are thus less likely to select similar objects or favour outliers as long as we have a reasonably diverse pool of candidate objects from which to sample.

This approach is predicated on being able to estimate the posterior label probability at least with reasonable accuracy, which is not guaranteed with standard machine learning methods. One of the approaches previously developed to better understand how confident one could be in a prediction is Conformal Prediction (CP), where prediction sets or ranges are formed corresponding to a prescribed confidence (Vovk et al., 2014). To derive a prediction for a new object a nonconformity score is assigned to it by applying some function, in most cases this is a prediction from a machine-learning (ML) model. This nonconformity score is then compared to nonconformity scores of other objects, where the label is

[1] the $\lbag\ \rbag$ notation denotes a bag or multiset, i.e. a collection in which elements can appear multiple times

known, by applying the same function. A randomness test is then performed that determines whether a particular label or region should be part of the prediction. In brief, define a non-conformity measure $A : \mathbf{Z} \times \mathbf{Z}^{n-1} \to \mathbb{R}$ which measures how different a given observation, z_i is to other observations in the set, $T_i = \wr z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n \int$. A general approach to forming a non-conformity measure is to use $1 - \hat{P}_\theta(y_i|x_i)$, with θ being trained on T_i and $\hat{P}(y|x)$ provided by the underlying ML method being the approximate posterior probability of observing label y for object x. Now let,

$$\alpha_i = A(z_i, T_i)$$

and assign a p-value to each possible label hypothesis, $y \in \mathbf{Y}$ for the object $x_{n+1} \in \mathbf{X}$:

$$p(y) = p((x_{n+1}, y), \wr z_1, \dots, z_n \int)$$

with,

$$p((x_i, y), T_i) = \frac{|\{j = 1, \dots, n : \alpha_j \geq \alpha_i\}|}{n}.$$

We use $\hat{y}_i = \text{argmax}_{y \in \mathbf{Y}} p((x_i, y), \dots)$ as the predicted label and denote the largest p-value, $p((x_i, \hat{y}_i), \dots)$ the credibility and the second largest, $1 - \max_{y \in \mathbf{Y} \setminus \{\hat{y}_i\}} p((x_i, y), \dots)$ the confidence. We should point out that the terminology used here may seem confusing as the *least confident* approach so-named since (Culotta and McCallum, 2005) using conformal prediction terminology could more appropriately be called the *least credible* approach.

The nice property with CPs is that if the data that is being modelled is generated randomly from some unknown probability distribution, then a CP will produce correct predictions for a fraction of the predictions it makes corresponding to at least the confidence. This property is called *validity* (Vovk et al., 2014) and seemingly provides a natural solution to our desire for calibration. It should be noted that CPs as formulated above are conservatively valid and that one can introduce a smoothing term (Vovk et al., 2014) to provide exact validity.

One drawback with CPs as formulated above is the computational cost required in forming the non-conformity measures α_i as each of these require retraining the underlying ML model. This, so-called transductive approach to conformal prediction can instead be replaced by an inductive approach (Vovk et al., 2014) alleviating the need for model retraining. Inductive conformal predictors (ICPs) split the set of training observations into a *proper training* set and a *calibration* set. The observations in the proper training set are then used to form the non-conformity measures for each member of the calibration set together with any new objects under consideration.

Finally, in active learning where we use the algorithm to select which new labels to open we clearly violate the underlying CP assumption of exchangeability since we no longer can view the objects as drawn at random. Originally proposed by (Zhou et al., 2013) and more recently adopted by (Nouretdinov, 2017) we can use a transfer learning approach in an inductive setting to retain validity. This leads to the algorithm shown in Alg. 1.

The trick here is to ensure that the *calibration* set is formed from the initial training set only since the generating distribution for the training set will deviate from that of the initial one as we actively transfer new samples for inclusion into the *proper training* set.

As mentioned, the calibration set C must be chosen from T_0 but as the overall training set grows can equally do so. Arbitrarily, we initially choose C at half the size of T_0 at random

Algorithm 1: Pool based active learning in batch mode with inductive conformal prediction

```
1. Require: A set of initial training observations, $T_0$ and a set of objects $S$, credibility threshold $\epsilon$, batch size $m$, and exploration budget $l$
2. Train ICP, $\chi_0$ on proper training set $T_0 \backslash C$ with calibration set $C \subset T_0$
3. For $k = 0, 1, \ldots, l$
    4. Form a set of candidates $Q = \lbag x | x \in S, \forall y \in \mathbf{Y} : p_{\chi_k}((x, y), C) \leq \epsilon \rbag$
    5. While $|Q| < m$: $Q = Q \uplus \lbag \mathrm{argmin}_{x \in S \backslash Q} \mathrm{argmax}_{y \in \mathbf{Y}} p_{\chi_k}((x, y), C) \rbag$
    6. Sample $m$ objects $\lbag x_1^k, \ldots, x_m^k \rbag$ without replacement from $Q$
    7. Open their labels $y_i^k$, $i = 1, \ldots, m$
    8. Update the training set $T_{k+1} = T_k \uplus \lbag (x_1^k, y_1^k), \ldots, (x_m^k, y_m^k) \rbag$
    9. Reselect calibration set, $C$ from $T_0$ s.t. $C \subset T_0$
    10. Train ICP, $\chi_{k+1}$ on proper training set $T_{k+1} \backslash C$ with calibration set $C$
11. Return $T_{l+1}$, and $\chi_{l+1}$
```

(under the constraint of having all classes present in the proper training set remaining). As the training set grows we linearly increase the calibration set size such that $|C| = \min(\max(|T_0|/2, |T_{k+1}|/5), |T_0|)$.

Whilst Alg. 1 operates with a preset exploration budget, l, an additional benefit of using CPs is that one can estimate the credibility on out of bag samples from the unexplored pool, $S \backslash T_k$ and alternatively use this to decide when adequate performance has been attained as a stopping criterion.

3. Results

3.1. Evaluation on UCI Covertype dataset

To confirm the expected behaviour on the proposed algorithm, Alg. 1 we investigate the performance on the Covertype dataset from UCI (Blackard and Dean, 1999). The dataset originates from a study of forest cover types in northern Colorado, US and has a number of cartographic variables as features alongside seven different forest cover type classes. To create a less complex, but imbalanced dataset we chose to reduce this by only considering the three cover types spruce (1), ponderosa pine (3) and aspen (5) leading to a dataset with a total of 257,087 observations with 54 features and a class density of approximately 82%, 14% and 4% respectively. From this we randomly sampled a validation set of 3000 objects across the three classes, leading to 2463, 425 and 112 observations for covertypes 1, 3 and 5 respectively.

In Fig. 1a we show the average performance of 100 repeats of random sampling, the classical least confident approach, query by transduction (QbT) (Balasubramanian et al., 2014) and the joint entropy based method by Nguyen et al. (Nguyen et al., 2012) on this

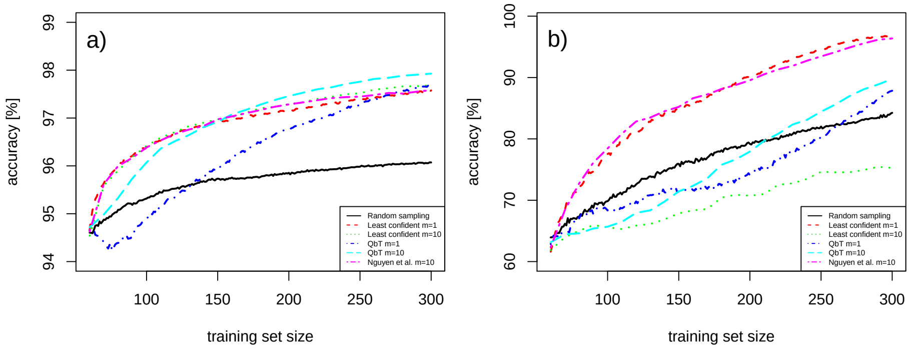

Figure 1: Performance of least confident approach, Query by Transduction (QbT) (Balasubramanian et al., 2014), and joint entropy based approach (Nguyen et al., 2012) compared to random sampling for the UCI Covertype dataset. Lines show average accuracy on the validation set over 100 repetitions. Subfigure a) is the original dataset, subfigure b) shows the modified dataset enriched with duplicates.

dataset. We use an approximately randomly selected initial training set (constrained to ensure all three classes are represented) and random forest (Breiman, 2001) as the underlying machine learning approach. We use $1 - \Delta_v$ as non-conformity measure with Δ_v being the proportion of votes in the ensemble (a surrogate for $P(y|x)$), i.e.,

$$A(z, T) = 1 - \Delta_v(z, T),\ \Delta_v((x, y), T) = |\{j = 1, \ldots, n : v(x, T)_j = y\}|/n$$

where $v(x, T)_j$ denotes the j'th vote on x in the ensemble of n random forest trees formed from T.

Given the size the dataset we also limit the exploration by sampling at random from S rather than evaluating all of S.

As expected one observes that in non-batch mode, $m = 1$ as well as in batch mode, $m = 10$ the least confident, QbT and the entropy based methods all have higher learning rates than random sampling but with little difference between the three.

To illustrate the problem with highly correlated observations in batch mode we next created an artificial dataset by subsampling from S, augmenting this with copies of the least confident objects, and replicating m times, with $m = 10$ as the batch size.

In Fig 1b we observe that the least confident approach in non-batch mode as well as the joint entropy based method both learn faster than random sampling. As expected, the least confident approach in batch mode learns much slower than random sampling (as we essentially keep selecting multiple copies of the same objects within each batch). QbT (irrespective of whether in batch mode or not) seems to initially learn slower than random sampling albeit eventually catching up.

Next we perform the same experiment using the ICP procedure defined in Alg. 1. Fig. 2a consistently show that the ICP algorithm performs as well as any of the other approaches irrespective of whether in a batch setting or not. In Fig. 2b the ICP algorithm in non-

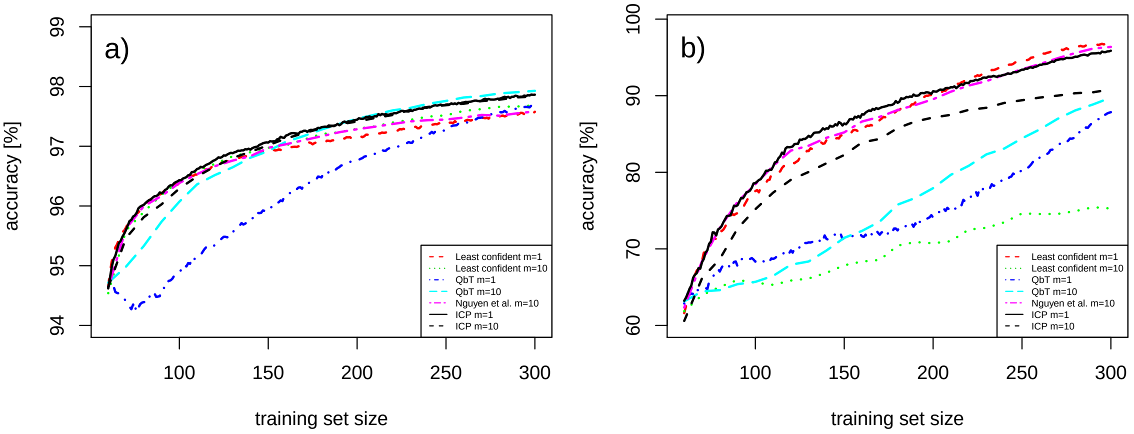

Figure 2: Performance of inductive conformal prediction (ICP) approach from Alg. 1 using $\epsilon = 0.10$ compared to the least confident, QbT and joint entropy based approaches. Lines show average accuracy on the validation set over 100 repetitions. Subfigure a) is the original dataset, subfigure b) shows the modified dataset enriched with duplicates.

batch mode performs as well as the least confident approach in non-batch mode or the joint entropy method, whereas it in batch mode learns a little slower, but still much faster than the remaining approaches.

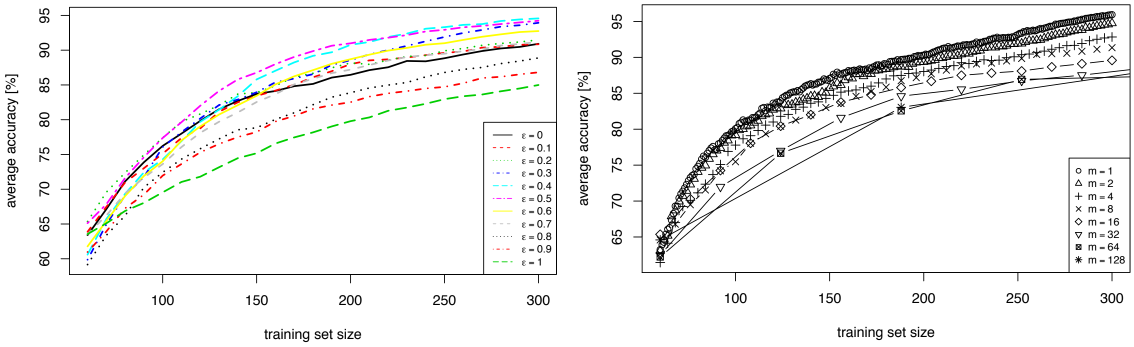

Figure 3: Average performance on the validation set over 100 repeats of the inductive conformal prediction approach from Alg. 1 on the modified dataset for varying ϵ and batch size, m. When varying ϵ the batch size is kept constant at 10 and when varying the batch size, ϵ is kept constant at 0.1.

In Fig. 2b we see that Alg. 1 in batch mode with $m = 10$ showed only slightly reduced performance compared with the non-batch equivalent. Next we evaluate the performance of Alg. 1 under varying batch size, m, and ϵ. Fig. 3 shows how increasing ϵ initially leads to improved performance up to a point (approximately 0.4) after which it declines. One also observes how increasing the batch size overall reduces performance.

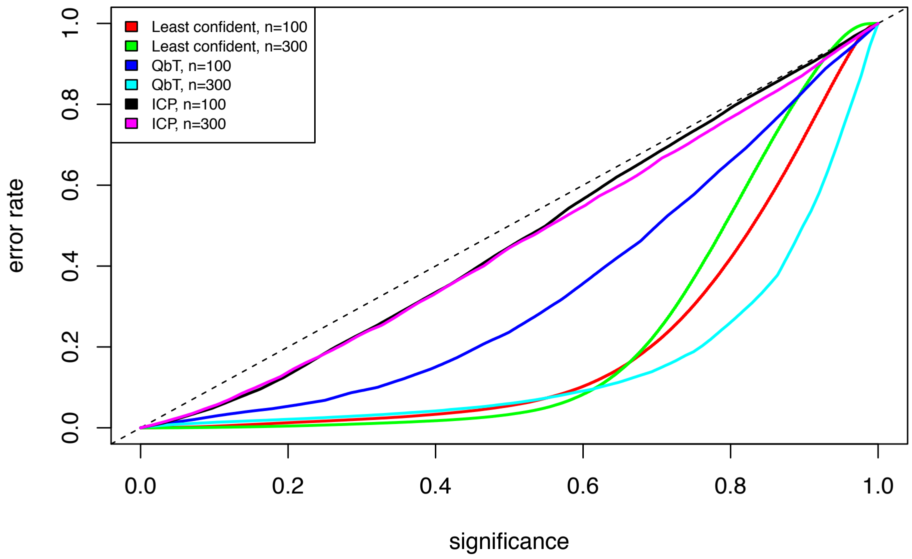

Figure 4: Calibration plot showing observed error rate on the validation set versus significance level over 100 repeats. The plots are formed at an early stage during active learning on the modified dataset (with a training set size, $n = 100$) as well as at the end ($n = 300$). In all cases, the batch size, m was 10.

Finally we assess the validity of the approaches experimentally. Fig. 4 shows how the random forest algorithm used in the least confident approach is overly conservative but with similar lack of calibration irrespective of whether we evalute this early on during active learning or at the end. QbT has some deviation from calibration initially and this increases during learning. The ICP approach remains approximately, conservatively valid irrespective of where it is assessed.

3.2. Evaluation on DDR dataset

We next applied the approach of Alg. 1 to the problem of phenotype classification in an image dataset. The dataset consists of imaging DNA damage response (DDR) in a cellular context under treatment and is relevant to understanding combination treatment in oncology in particular.

The dataset consists of a large amount of data for single cells, captured via confocal microscopy. Briefly, cells were engineered to express a fluorescent tag onto histones to observe dynamic changes in DNA structure, or chromatin. These cells were imaged using a spinning disk confocal microscope (CV7000, Yokogawa) with a 60X objective for 48 hours, with an image captured every 6 minutes. From the images, individual cells were identified and 99 textural and morphological features were calculated for each cell (Columbus software, Perkin Elmer).

Expert annotation is performed by direct visualisation of individual cell images. There are seven primary classes or phenotypes in this dataset, which are illustrated in Fig 5. Due to the biology of the system, the phenotypes are highly unbalanced; cell division, or mitosis, is usually fast compared with the rest of the cell cycle, and when cell death occurs, the cells often detach from the surface and disappear from the image. Therefore, these classes are comparatively rare in the dataset. The aim of the experiment is to identify abnormalities in the cell division process caused by drug treatment, and therefore these abnormalities could form an even smaller subset of mitotic cells. We include an unknown class, which is used when it is not possible to classify a cell into one of the primary phenotypes — this can be due to an unusual cellular appearance, but more likely due to a faint or out of focus image which prevents accurate classification.

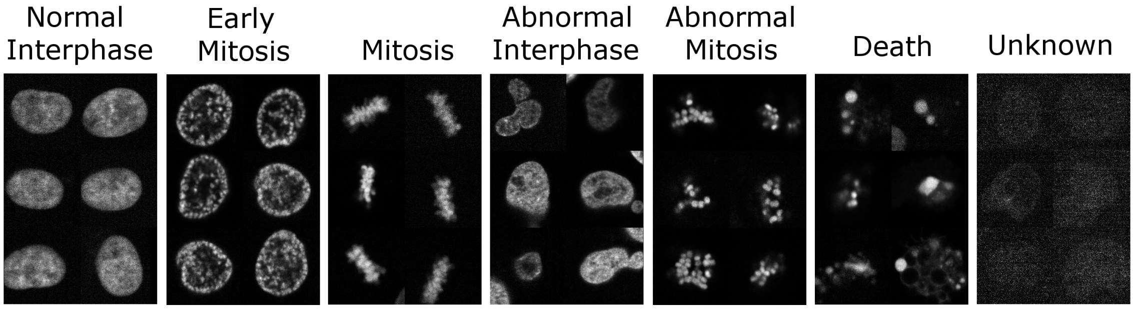

Figure 5: Characteristic examples of the primary phenotypes observed in the DDR imaging dataset.

The workflow for our active learner is as follows: a small initial training set of 200 samples was constructed by direct browsing and selection from the cell images by an expert. Based off this initial set, Alg. 1 was used to select a batch of 16 unlabelled images from the remaining pool, which were then presented to the expert to be annotated. These samples were then added to the training set and the process repeated for a total of 60 iterations. Fig 6 shows a representative iteration, with the batch having been annotated via colour-coded borders. Due to the nature of the dataset, we do not have access to a representative validation set to assess the accuracy of the model as training progresses; instead, we illustrate the performance of our active learner through the average Out of Bag credibility across a random sample of images (Fig 7), which shows an increase over the initial manual selection, before tending towards a plateau around 90%.

We next illustrate the use of our trained model to study the effect of compound treatment on the mechanism of cell death at different time points. From 12 hours onwards, the compound treatment shows an increase in the proportion of abnormal cells, both in interphase and mitosis, as well increasing levels of cell death (Fig 8). The relative levels of abnormalities shed light on whether a compound is having its effect during interphase

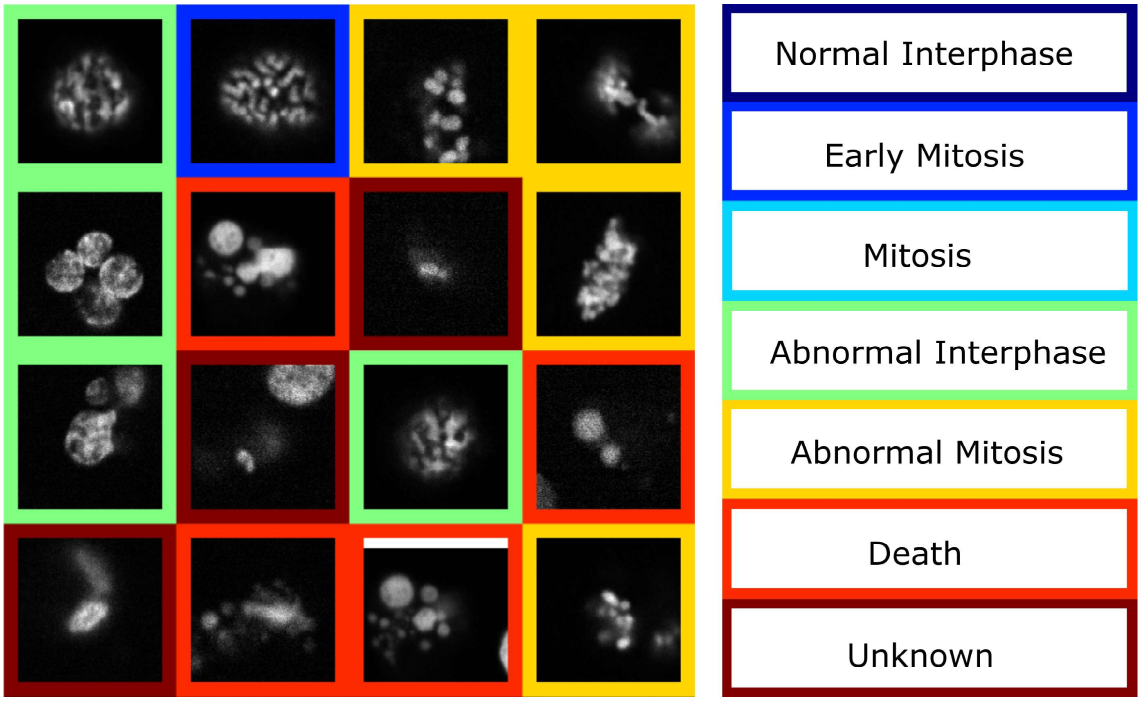

Figure 6: Illustration of the interface for batch annotation. The expert is presented with the selected batch of cellular images from the unlabelled pool. These are then manually annotated, with colour-coding to highlight the assigned class. The image shows a representative iteration, typically consisting of interesting cellular appearances, and without duplication or redundancy in the selected cells.

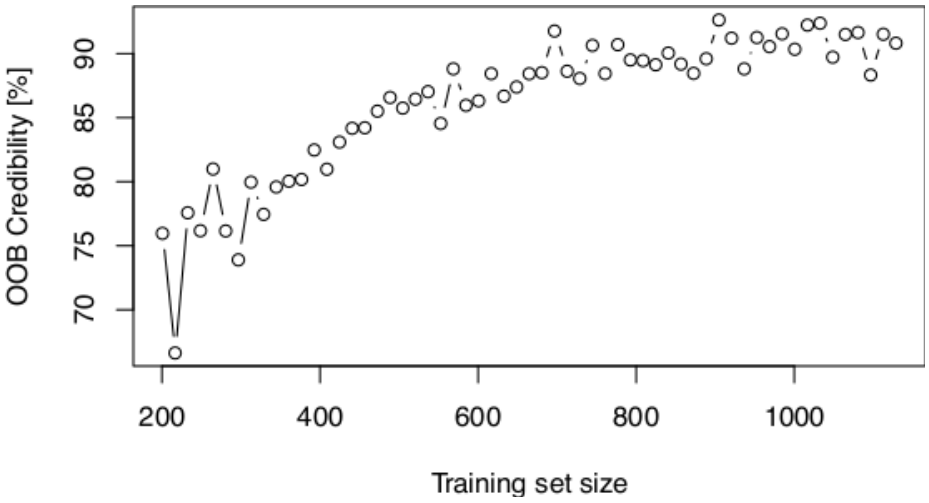

Figure 7: Performance of the ICP approach from Alg. 1 using $\epsilon = 0.30$ on out of bag samples drawn from the unexplored pool. In the absence of true validation data the average credibility returned by the ICP is presented as a surrogate of predictive performance vs increase in training set size.

or mitosis, and whether this causes cells to enter apoptosis (cell death), continue through the cell cycle with abnormalities, or remain in interphase without dividing. One notable observation from Fig 8 is the increase in the 'Unknown' class over time, with a clearly larger fraction of control cells classified in this way, which is caused by two effects. Due to the nature of live cell imaging, the fluorescent tag generally suffers photobleaching over time, leading to fainter images which are less likely to be able to be classified into one of the biological phenotypes. Secondly, through cell divisions the density of cells increases over time. The proximity of neighbouring cells causes some cells to be pushed out of the microscope focus, giving a blurred appearance which is classified as unknown. This also explains the increase in unknowns in control cells compared to compound-treated, as there is less death and more cell division in control cells, leading to higher cell density. This highlights the importance of including an 'Unknown' class to prevent differential densities affecting the accuracy of classification.

4. Discussion

In the above we have shown how active learning approaches can fail to learn well in batch mode unless specifically designed to avoid selecting instances that are highly correlated. Often a naïve approach when extending an active learner to batch mode is simply to select the top most relevant examples within each batch based on some, learner specific, measure of relevance. Here we instead suggest to randomly select from the pool of examples considered *sufficiently* relevant. What constitutes *sufficient* can be problem or method specific but the

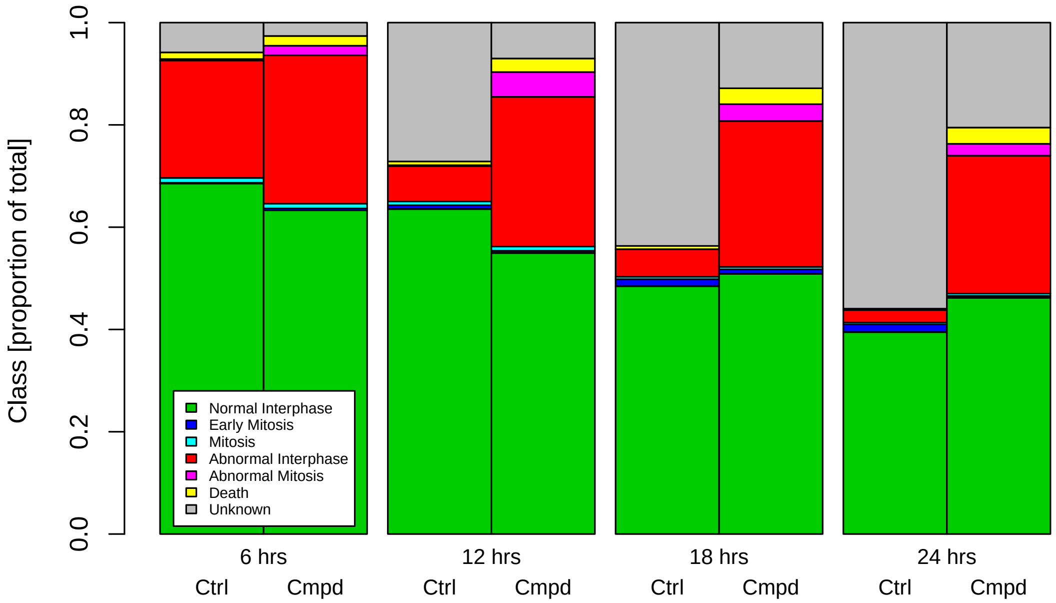

Figure 8: Comparison of control (ctrl) versus treatment with compound (cmpd) on the cell populations over time.

use of conformal prediction can provide a natural interpretation of relevance because of its inherent validity guarantees.

Considering relevance, for the ICP based approach (Alg. 1), we observed in Fig. 3 how a too small ϵ prohibited sufficient diversity in the selected examples whereas a too large ϵ implied insufficient prioritisation amongst the examples. All in all, one would expect the optimal choice of ϵ to be problem specific and one will need to strike a balance between using domain understanding with that of treating it as yet another hyperparameter to select.

Having a method which is well calibrated does however not only improve the ability to interpret a given ϵ but equally may inform when to stop exploration. In Fig. 7 we observed how credibility seemed to plateau around 90% suggesting limited benefit of opening additional labels beyond this point. If however, the method does not remain valid (or approximately so) as we observed for QbT (which irrespective of using the conformal framework clearly violates the assumption of exchangeability and thus fails to retain calibration) one cannot use such estimates to make decisions to halt exploration with confidence.

When the classes are strongly imbalanced, a potential improvement might be obtained by integrating Mondrian ICP into the batch mode active learning framework to ensure validity on a per-class basis and improve performance for the rare classes.

4.1. Discussion on DDR dataset

In the typical cellular imaging workflow in current use, cells are annotated with phenotype information and added to the training set concurrently with the set up of the image analysis pipeline, which involves manual selection of a field of view and timepoint in the experiment, followed by manual choice of which cells in the field of view to annotate and add to the training set. There is potential to make considerable improvements to this process through an active learning approach.

There are some practical considerations to be taken into account when applying our approach to full cellular imaging datasets, which can typically consist of several million unlabelled cells. The adoption of ICPs provided a considerable improvement in the speed of computation, and by subsampling 10^5 samples from the unexplored pool at each iteration, we were able to implement this approach in an online setting, where the expert annotates the chosen samples in batches of 16 at a time, and is able to run through up to 60 iterations in a single sitting. Here it is important to balance the improved performance from using small batch sizes, with the efficiency of using the time of the expert to annotate in an online setting. As described above, the choice of ϵ is problem-specific; the nature of the time-lapse DDR dataset, with sequential data points likely to be highly correlated, meant that we used a relatively large value of ϵ in order to ensure diversity in the examples opened for annotation.

In oncology, targeting of the DDR has increased exponentially in the past few years. Most of these agents elicit a cell cycle perturbation, which remains poorly characterised. The clear majority of DDR inhibitors are used in combination therapies with either standard of care (i.e. radiation or chemotherapy) or with themselves. Understanding molecular determinants of therapeutic response in DDR is of paramount importance, particularly in combination therapies where doses are often sub-optimal due to lack of biomarkers. Quantifying phenotypic response and faithfully recapitulating heterogeneity of response provide deeper understanding of single agent vs combination response (Gascoigne and Taylor, 2008; Tyson et al., 2011), which could be exploited for optimal dose predictions. The approach has also the potential to underpin undesired side effects arising from genotoxic stress and wider genomic instability having a wide potential for translational impact. In the DDR imaging space, the main phenotypes of interest are rare, as the processes of cell division and cell death are very fast in comparison with the overall cell cycle. Based on the appearance of the chromatin, which is fluorescently marked in our images, it is possible to discern the mechanism by which a drug or treatment is having an action. This necessitates expert input in order to differentiate between subtly different cellular appearances, which are rare events in the dataset. A manual search of the images to find and classify the phenotypes of interest is therefore highly inefficient.

There are several additional drawbacks to delegating the choice of training samples entirely to the end user. If selection is done in an ad hoc manner, the data is likely to be both unbalanced, as rare phenotypes, such as death and abnornal mitosis in the DDR dataset are much harder to find, and highly redundant, as chosen examples will contain overlapping information. A further shortcoming of manual selection is that there is no mechanism to ensure that the chosen examples are fully representative of each class - highly typical examples will be overrepresented, simply because they are easier to classify.

Our approach solves these problems by selecting those examples from the pool which are perceived as novel. In doing so, our active learner finds examples which are not well represented in the existing training data. For DDR cell images, this corresponds to finding interesting cells with atypical appearance, for which the model does not make confident predictions. With modification, our approach could be used as a phenotype discovery method (Naik et al., 2016), to identify and annotate novel classes. This has applications in phenotypic screening, where the effect of a compound or treatment is observed through changes in the cellullar phenotype. Although a small number of classes are known in advance through the use of positive and negative controls, we wish to identify new phenotypes within the full dataset corresponding to different mechanisms of action that were not known *a priori*. Active learning via conformal predictors is a powerful technique to efficiently identify those cells which do not resemble the control data, making better use of the expert's time when annotating training data, and providing a new workflow for phenotypic screening where all the phenotypes do not need to be known from the controls.

5. Summary

We have designed and implemented an active learning framework, based around conformal predictors, and demonstrated its effectiveness on the UCI Covertype dataset. When applied to a cellular image dataset with a number of subtly different phenotypes, we have shown this approach enables rapid generation of a representative training set. By selecting objects for annotation based on perceived novelty, we are able to efficiently identify rare phenotypes within the dataset, which is important for categorising the mechanism of drug action, in DNA damage response studies. This new workflow enables unbiased and efficient phenotype detection, which will greatly enhance the quantitative information that can be extracted from complex cellular screening data.

Acknowledgments

We acknowledge many productive conversations with colleagues at the Royal Holloway University of London, in particular Paolo Toccaceli, Alex Gammerman and Vladimir Vovk.

6. References

References

Vineeth N Balasubramanian, Shayok Chakraborty, Shen-Shyang Ho, Harry Wechsler, and Sethuraman Panchanathan. *Conformal Prediction for Reliable Machine Learning*. Elsevier Inc., 2014.

Jock A Blackard and Denis J Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 24(3):131–151, 1999.

Leo Breiman. Random Forests. *Machine learning*, 45(1):5–32, 2001.

A Culotta and A McCallum. Reducing labeling effort for structured prediction tasks. *AAAI*, 2005.

Karen E. Gascoigne and Stephen S. Taylor. Cancer cells display profound intra- and interline variation following prolonged exposure to antimitotic drugs. *Cancer Cell*, 14(2):111–122, 2008.

Yuhong Guo and Dale Schuurmans. Discriminative batch mode active learning. *Advances in neural information processing systems*, 179:593–600, 2008.

Steven C H Hoi, Rong Jin, Jianke Zhu, and Michael R Lyu. *Batch mode active learning and its application to medical image classification*. ACM, New York, New York, USA, June 2006.

Sergio Matiz and Kenneth E Barner. Inductive conformal predictor for convolutional neural networks: Applications to active learning for image classification. *Pattern Recognition*, 90:172–182, June 2019.

A W Naik, J D Kangas, D P Sullivan, and R F Murphy. Active machine learning-driven experimentation to determine compound effects on protein patterns. *eLife*, 2016.

Hieu T Nguyen, Joseph Yadegar, Bailey Kong, and Hai Wei. Efficient batch-mode active learning of random forest. *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pages 596–599, May 2012.

Ilia Nouretdinov. Reverse conformal approach for on-line experimental design. *Proceedings of Machine Learning Research*, 60:1–8, 2017.

Sachin Ravi and Hugo Larochelle. Meta-Learning for Batch Mode Active Learning. In *ICLR*, February 2018.

N Roy and A McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Int. Conf. on Machine Learning*, 2001.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active Hidden Markov Models for Information Extraction. In *Advances in Intelligent Data Analysis*, pages 309–318. Springer, Berlin, Heidelberg, Berlin, Heidelberg, September 2001.

Burr Settles. Active Learning Literature Survey. Technical Report 1648, University of Wisconsin - Madison, 2009.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-Instance Active Learning. In *Neural Information Processing Systems NIPS*, pages 1289–1296, 2008.

C E Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948.

John J. Tyson, William T. Baumann, Chun Chen, Anael Verdugo, Iman Tavassoly, Yue Wang, Louis M. Weiner, and Robert Clarke. Dynamic modelling of oestrogen signalling and cell fate in breast cancer cells. *Nature Reviews Cancer*, 11:523, 2011.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer-Verlag Berlin, Heidelberg, December 2014.

S Zhou, E N Smirnov, H B Ammar, and R Peeters. Conformity-Based Transfer AdaBoost Algorithm. *IFIP International Conference ...*, 2013.