Applying Conformal Prediction to the Bovine TB Diagnosing

Dmitry Adamskiy[2], Ilia Nouretdinov[2], Andy Mitchell[1], Nick Coldham[1], and Alex Gammerman[2]

[1] Veterinary Laboratories Agency,
{a.p.mitchell,n.g.coldham}@vla.defra.gsi.gov.uk
[2] Royal Holloway, University of London,
{adamskiy,ilia,alex}@cs.rhul.ac.uk

Abstract. Conformal prediction is a recently developed flexible method which allows making valid predictions based on almost any underlying classification or regression algorithm. In this paper, conformal prediction technique is applied to the problem of diagnosing Bovine Tuberculosis. Specifically, we apply Nearest-Neighbours Conformal Predictor to the VETNET database in an attempt to allow the increase of the positive prediction rate of the existing Skin Test. Conformal prediction framework allows us to do so while controlling the risk of misclassifying true positives.

Keywords: conformal predition, bovine TB, online learning

1 Introduction

Bovine Tuberculosis (bTB) is an infectious disease of cattle, caused by the bacterium Mycobacterium bovis (M.bovis). The disease is widespread in certain areas of the UK (particularily South West England) and of major economic importance, costing the UK Government millions of pounds each year, since positive animals are slaughtered and compensation paid to the cattle owners. The main testing tool for diagnosing TB in cows is the Single Intradermal Cervical Tuberculin (SICCT) skin test. The procedure involves administering intradermally both Bovine and Avian Tuberculin PPDs and measuring the thickening of the skin.

Avian tuberculin is used to exclude unspecific reactions so the actual value of the skin test is a difference of thickenings: $(B_2 - B_1) - (A_2 - A_1)$, where A_1 and B_1 are the initial values of skin thickness measured in millimetres, A_2 and B_2 are skin thickness after the injection of avian and bovine tuberculin respectively.

If the cow is a reactor, in the sense that the test is positive,

$$(B_2 - B_1) - (A_2 - A_1) > T$$

where T is a threshold (usually 3mm), it is slaughtered and the post-mortem examination is performed, which may result in the detection of visible lesions typical of M.bovis. Furthermore, samples for some of the slaughtered cattle are sent for the bacteriological culture analysis.

The data on all the reactors is stored in the VETNET database. The data stored there includes (per reactor) the numeric test results A_1, A_2, B_1, B_2, and such features as age, herd identifier, date of the test, post-mortem and culture results (if any) and others([5]).

As there are two tests that can confirm the diagnosis after the cow is slaughtered, the definition of truly positive cow could be different and here we use logical OR as such (the cow is positive if there are visible lesions or if the culture test was positive). The data in VETNET alone is not enough to judge about the efficiency of the skin test, as the post-mortem tests are not performed for the negative animals. However, it is believed that the positive prediction rate could be improved by taking into account some other factor apart from just the binary result of the skin test.

In this paper we state it as an online learning problem: given the history of the animals tested prior to the current reactor, we try to dislodge, at a given significance level, the hypothesis of this reactor being a real one. In what follows we introduce conformal predictors and describe how conformal predictors could be used for this task.

2 Conformal prediction

Conformal prediction [1] is a way of making valid hedged predictions which does not require any assumption other than i.i.d.; the only assumption made is the i.i.d. assumption: the examples are generated from the same probability distribution independently of each other.

Also, it is possible to estimate confidence in the prediction of the given individual example. Detailed explanation of conformal prediction could be found in [1], here we outline the intuition behind it and the way it is applied to the problem stated.

Online setting implies that the sequence of examples $z_i = (x_i, y_i)$, $z_i \in Z = (X, Y)$ is revealed one by one and at each step after the object x_i is revealed the prediction is made. Original conformal prediction algorithm takes as a parameter the function called nonconformity measure and outputs the prediction set Γ. Thus the performance of the conformal predictor is measured in terms of validity and efficiency: how many errors the predictor makes ($y_i \notin \Gamma_i$) and how big is the set Γ_i. In case of the classification task it is also possible to output forced point prediction along with the measure of confidence in it.

A nonconformity measure formally is any measurable function taking bag of examples and a new example and returning a number specifying "nonconformity" (strangeness) of a given example to the set. The resulting conformal predictor will be valid no matter what function we choose, however in order to obtain an efficient predictor one should carefully select the reasonable one. Specifically there is a general scheme of defining nonconformity from any given point predictor.

Algorithm 1 Conformal Predictor for classification

```
**Input:** data examples $(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l) \in X \times Y$
**Input:** a new object $x_{l+1} \in X$
**Input:** a non-conformity measure $A : (z_i, \{z_1, \ldots, z_{l+1}\}) \to \alpha_i$ on pairs $z_i \in X \times Y$
**Input(optional):** a significance level $\gamma$
$z_1 = (x_1, y_1), \ldots, z_l = (x_l, y_l)$
**for** $y \in Y$ **do**
  $z_{l+1} = (x_{l+1}, y)$
  **for** $j$ in $1, 2, \ldots, l+1$ **do**
    $\alpha_j = A(z_j, \{z_1, \ldots, z_l, z_{l+1}\})$
  **end for**
  $p(y) = \frac{\#\{j=1,\ldots,l+1 : \alpha_j \geq \alpha_{l+1}\}}{l+1}$
**end for**
**Output(optional):** prediction set $R^{\gamma}_{l+1} = \{y : p(y) \geq 1 - \gamma\}$
**Output:** forced prediction $\hat{y}_{l+1} = \arg\max_y \{p(y)\}$
**Output:** confidence
$conf(\hat{y}_{l+1}) = 1 - \max_{y \neq \hat{y}_{l+1}} \{p(y)\}$
```

2.1 Mondrian conformal predictors

The algorithm 1 is valid in a sense that under the i.i.d. assuption it makes errors independently on each trial with probability less then $1 - \gamma$. However, sometimes we want to define the categories of the examples to have the category-wise validity. For instance, suppose that the examples fall into "easy to predict" and "hard to predict categories", then the overall validity will be reached by conformal predictor, but the individual error rate for "hard to predict" objects could be worse.

In order to overcome this, Mondrian conformal predictor (first presented in [2]) is used. Here we are interested in label-wise validity, thus we predict it in a most simple form, see Algorithm 2.

3 Applying conformal prediction to the VETNET database

We used the positively test cows from VEBUS subset of the original VETNET database. It includes 12873 false positives and 18673 true positives. In what follows the words "true positives" and "false positives" will refer to the skin test results.

After the preliminary study, it was discovered that the most relevant atrtibutes for classification are numeric value of skin test result, age and either the ID of the given test which is a herd identifier combined with a test date, or just indentifier of a herd (that may cover several tests performed at different time).

The extract from the VETNET database showing those features is shown in Table 1.

Algorithm 2 Mondrian Conformal Predictor for classification

```
**Input:** data examples $(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l) \in X \times Y$
**Input:** a new object $x_{l+1} \in X$
**Input:** a non-conformity measure $A : (z_i, \{z_1, \ldots, z_{l+1}\}) \to \alpha_i$ on pairs $z_i \in X \times Y$
**Input(optional):** a significance level $\gamma$
$z_1 = (x_1, y_1), \ldots, z_l = (x_l, y_l)$
for $y \in Y$ do
    $z_{l+1} = (x_{l+1}, y)$
    for $j$ in $1, 2, \ldots, l+1$ do
        $\alpha_j = A(z_j, \{z_1, \ldots, z_l, z_{l+1}\})$
    end for
    $p(y) = \frac{\#\{j=1,\ldots,l+1 : y_j = y, \alpha_j \geq \alpha_{l+1}\}}{|y = y_j|}$
end for
**Output(optional):** prediction set $R^{\gamma}_{l+1} = \{y : p(y) \geq 1 - \gamma\}$
**Output:** forced prediction $\hat{y}_{l+1} = \arg\max_y \{p(y)\}$
**Output:** confidence
$conf(\hat{y}_{l+1}) = 1 - \max_{y \neq \hat{y}_{l+1}} \{p(y)\}$
```

Table 1. Two entries in VETNET database

Test time (seconds)	CPHH (herd ID)	Lesions	Culture	Age	AvRes	BovRes
1188860400	35021001701	1	1	61	4	22
1218495600	37079003702	0	0	68	1	9

The first of these examples is a True Positive (Lesions or Culture test is positive) and the second is a False Positive (both Lesions and Culture tests are negative).

The task is to distinguish between these two classes.

Remind that the goal is to decrease number of cows being slaughtered. This means that we wish to discover as many cows as possible to be False Positives. On the other hand, the number of True Positives misclassified as False Positives should be strictly limited. So unlike standard conformal prediction, the role of two classes is different.

Thus we present a one-sided version of conformal predictor. Each new example is assigned only one p-value, that corresponds to True Positive hypothesis. Then for a selected significance level γ we mark a cow as a False Positive if $p < \gamma$. This allows us to set the level at which we tolerate the marking of true positive and we aim to mark as many false positives as possible.

The property of validity is interpreted in the following way: if a cow is True Positive, it is mismarked as a False Positive with probability at most γ. A trivial way to achieve this is to mark any cow with probability γ. So the result of conformal prediction can be considered as efficient only if the percentage of marked False Positives is essentially larger.

In our experiments we used the nonconformity measure presented in algorithm 3 based on k-Nearest-Neighbour algorithm.

Algorithm 3 kNN Nonconromity Measure for VETNET database

```
**Input:** a bag of data examples $z_1 = (x_1, y_1), z_2 = (x_2, y_2), \ldots, z_{l+1} = (x_{l+1}, y_{l+1}) \in X \times Y$
**Input:** an example $z_i = (x_i, y_i)$ from this bag;
**Input:** a distance function $d(x_1, x_2) : X \times X \to \mathbb{R}^+$
$A(z_i, \lbag z_1, z_2, \ldots, z_{l+1} \rbag) = |\{j : x_j$ is amongst $k$ nearest neighbours of $x_i$ in $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{l+1}$ according to the distance $d$, and $y_j \neq y_{l+1}\}|$
```

A possible version of efficient predictor can be done setting $k = 50$ and using the following distance, which assings the highest importance to comparison of herd IDs, second priority is given to the numerical value $(B_2 - B_1) - (A_2 - A_1)$ of skin test, and age is used as an additional source of information.

$$dist(x_1, x_2) = 100S(x_1, x_2) + T(x_1, x_2) + |log_{10}(age(x_1)) - log_{10}(age(x_2))|$$

where $S(x_1, x_2) = 1$ if x_1 and x_2 belong to the same herd and 0 otherwise, $T(x_1, x_2)$ is the difference between numerical values of test results on x_1 and x_2. Thus, first all the animals within given test are considered as neighbours and then all the others. Experiments showed that this distance resulted in the efficient predictor though the validity property holds for other parameters as well. The results on the subset of VETNET database are summarized in the Table 2. The experiment was performed in an online mode with the data sorted by test date (as in real life).

Table 2. VETNET results

Significance level	Marked reactors within FP	Marked reactors within TP
1%	1267/12873	138/18673
5%	4880/12873	919/18673
10%	7971/12873	1904/18673

The i.i.d. assumption is clearly a simplification here, but as in some other conformal predictor applications(see [4]) we can see that it is not essentially broken and we can see that the validity property holds: the number of marked reactors within true positives is indeed the level that was set. The efficiency could be judged by the number of marked reactors within false positives: at a cost of misclassifying 10% of true positives it is possible to identify almost two thirds of test mistakes.

4 Conclusions and Future work

We can see from the table above that the resulting predictions are valid and efficient. The disadvantage is not taking into account the delay (normally several weeks), needed to perform post mortem analysis. This actually might lead to overestimation of the importance of herd ID as a factor: when a skin test is performed on many cows from same farm same day, it is likely to be either correct or wrong on the most of them.

To perform the experiment more fairly, the online protocol can be replaced with "slow learning" one (described in [3]) where the labels are revealed not immediately, but with the delay. Preliminaty investigation show that herd ID in such case should be replaced with more specific attributes related to the illness history of a herd.

5 Acknowledgements

This work was supported in part by funding from the ERASySBio+ Salmonella Host Interactions PRoject European Consortium (SHIPREC) grant; VLA of DEFRA grant on Development and Application of Machine Learning Algorithms for the Analysis of Complex Veterinary Data Sets; EU FP7 grant O-PTM-Biomarkers (2008–2011); and by grant PLHRO/0506/22 (Development of New Conformal Prediction Methods with Applications in Medical Diagnosis) from the Cyprus Research Promotion Foundation.

References

1. V.Vovk, A.Gammerman, G.Shafer "Algorithmic Learning in a Random World", Springer, 2005.
2. Vladimir Vovk, David Lindsay, Ilia Nouretdinov, Alex Gammerman "Modrian Confidence Machine", Working Paper #4, 2003.
3. Daniil Ryabko, Vladimir Vovk and Alex Gammerman "Online Region Prediction with Real Teachers", Working Paper #7, 2003
4. Ilia Nouretdinov, Brian Burford, and Alex Gammerman "Application of Inductive Confidence Machine to ICMLA Competition Data" Proc. ICMLA, 2009, pp.435-438
5. Andy Mitchell, Rachael Johnson. Private communications.