Application of conformal prediction in a more formal definition of applicability domain

UK QSAR/MGMS meeting @ Cardiff – 11/04/2018

Sébastien Guesné
Scientist
sebastien.guesne@lhasalimited.org
Acknowledgements

• Thierry Hanser
  • Research Leader – Cheminformatics and Computer Science

• Lilia Fisk
  • Senior Scientist – Leads the science on Hepatotoxicity in the AOP* framework

• Richard Williams
  • Chief Scientific Officer – Leads the science at Lhasa Limited

*Adverse Outcome Pathway
Overview of the presentation

• A more formal definition of applicability domain
  
  • Context
  
  • Applicability domain and its challenges
  
  • Introduction to Lhasa Limited’s framework – applicability, reliability and decidability domains

• Conformal prediction
  
  • Introduction to Conformal Prediction’s framework
  
  • Schematic representation of Conformal Prediction’s output

• Application of Conformal prediction to the decidability level of Lhasa Limited’s framework

2 - Norinder U, Rybacka A, Andersson PL. Conformal prediction to define applicability domain - A case study on predicting ER and AR binding. SAR QSAR Environ Res. 2016, 27, 303-316.
A more formal definition of applicability domain

Hanser T, Barber C, Marchaland JF, Werner S

Applicability domain: towards a more formal definition.
• Risk assessment of a specific chemical structure
  • One cannot rely on the overall statistical performance of a model
  • “How much can I trust the model’s prediction for this specific compound?” regardless of its statistical performance
Applicability domain

• OECD’s definition¹
  • “The applicability domain of a (Q)SAR model is the response and chemical structure space in which the model makes predictions with a given reliability”

• The challenge
  • Applicability domain is a complex concept – it must account for:
    • Interpolation within the chemical structure space
    • Density of knowledge around the query chemical structure
    • Distance of the query chemical compound to the decision boundary of an in silico model

1 – OECD Guidance document on the validation of (quantitative)structure-activity relationship (Q)SAR models
Lhasa Limited’s Framework*

Applicability Domain

• Interpolation within the chemical structure space
Reliability Domain

• Density of knowledge around the query chemical structure

*adapted from Aniceto N, Freitas AA, Bender A, Ghafourian T A novel applicability domain technique for mapping predictive reliability across the chemical space of a QSAR: reliability-density neighbourhood Journal of Cheminformatics, 2016, 8, 69.
Decidability Domain

- Distance of the query chemical compound to the decision boundary of an *in silico* model
Conformal prediction (CP)

Norinder U, Rybacka A, Andersson PL. *Conformal prediction to define applicability domain - A case study on predicting ER and AR binding*. SAR QSAR Environ Res. 2016, 27, 303-316.

CP is a broad concept

• Today - I am going to discuss a specific CP
  • Inductive CP for binary classification problems: label = active or inactive
  • Mondrian CP – ensure validity on the two labels
  • Random Forest as the underlying MLA

• I will not discuss
  • Transductive CP
  • CP for regression problems
  • Other types of underlying MLA (e.g. Support Vector Machine, K-Nearest Neighbours)
Important considerations on CP

• CP is an algorithmic framework

  • CP needs an underlying machine learning algorithm (MLA)

  • CP needs a non conformity function – defined by the user

  • CP may change the prediction of the underlying MLA - calibration

  • CP and an underlying MLA produce predictions complemented with the information on their “reliability” – $p$-value
Properties of Conformal Predictors

• Validity
  • Always valid if data are independent and identically distributed (i.d.d.)
  • At a chosen significant level $\varepsilon$ (0-100%), the error rate of CP for each label never exceeds the significant level

• Efficiency
  • The output of CP contains as few as possible multiple predictions
  = Is the conformal prediction informative?
Framework of CP

Training Set

30%

70%

Calibration Training Set

Proper Training Set

Random Forest

Inductive Mondrian CP
Binary classification problem
Random Forest

*Number of trees predicting the class divided by the total number of trees.
Nonconformity scores $\beta$

- $\beta = 1$ means maximum unconformity
- **Calibration set**
  - If label = active then $\beta_{\text{act}} = 1 - P(\text{active})$ from prediction
  - If label = inactive then $\beta_{\text{inact}} = 1 - P(\text{inactive})$ from prediction
- **Query set**
  - $\beta_{\text{act}} = 1 - P(\text{active})$ from prediction
  - $\beta_{\text{inact}} = 1 - P(\text{inactive})$ from prediction

Non conformity function

Inductive Mondrian CP
Binary classification problem
Random Forest
**p-values calculation of queries**

- 2 list of βs from the calibration set
- 2 βs per query

\[
p\text{-value} = \frac{43 + U(0,1)}{77 + 1} = 0.557
\]

\[
p\text{-value} = \frac{6 + U(0,1)}{70 + 1} = 0.096
\]

Inductive Mondrian CP
Binary classification problem
Random Forest

\[
\beta
\]

\[
\text{conform}
\]

\[
\text{nonconform}
\]
The output of Mondrian CP

- One \( p \)-value per label that defines 3 regions
  - “Both” region = both labels, active and inactive, are predicted
  - “Unique label” region = label of the highest \( p \)-value is predicted (active or inactive)
  - “Null” region = no label is predicted
- The significant level \( \varepsilon \) (0-100%) determines which region is predicted by CP – \( \varepsilon \) is defined by the user
Schematic representation of CP’s output

- p-value (Inactive) = 0.096
- p-value (Active) = 0.557

“Both” Region = Active And Inactive

“Unique label” Region = Active

“null” Region

ε (significant level)

confidence = 1 − ε

Inductive Mondrian CP
Binary classification problem
Random Forest
Generalisation – 2 labels

Inductive Mondrian CP
Binary classification problem
Random Forest

1 - \( \varepsilon \) = confidence
Generalisation – 3 labels

“Three labels” Region = label 1, 2 and 3

“Two labels” Region = label 2 and 3

“Unique label” Region = Label 3

“null” Region

$\varepsilon$ (significant level)
Application of Conformal Prediction to the decidability level of Lhasa Limited’s framework
Dataset and MLA

• Dataset
  • BSEP inhibitors
  • 610 structures
  • 321 active and 289 inactive compounds

• Machine Learning Algorithm
  • Random Forest
  • Equal size sampling with replacement
  • 1000 trees
  • Descriptor sampling: square root
  • Split selection: Gini index

• Assumption
  • All chemical structures belong to the Applicability and Reliability Domains of the model as defined in the first two steps of the Lhasa Limited’s framework
Validity of the CP
Decidability – 1\textsuperscript{st} approach

- Confidence = Decidability
- Prediction at that level of confidence = \(1 - \varepsilon\)

\[\varepsilon\text{ (significant level)}\]

“Both” Region = Active And Inactive

“Unique label” Region = Active

“null” Region
Results

Regions of the Conformal Prediction

- "Both" Region
- "Unique label" Region
- "null" region

1 - \( \varepsilon \) = confidence
Decidability – 2nd approach

- Decidability = $|\Delta p\text{-value}|$
- Prediction label of the highest p-value

Active with a decidability of 46.1%
Results - correct vs incorrect prediction
Results – Balanced Accuracy vs Decidability

One instance of the training set

1000 bootstrap samples of the training set
Results – Balanced Accuracy vs Coverage

One instance of the training set

1000 bootstrap samples of the training set
Conclusion

• Lhasa Limited’s framework of applicability domain
  • Based on three independent stages
  • This separation allows each step to be clarified and formalised
• Second approach for the decidability domain is reasonable
  • Expected behaviour
  • Calibration of the random forest output
• Future work
  • Will be applied to Lhasa Limited’s technologies
  • Design of a non conformity function within Lhasa Limited’s technologies
Thank you for listening!
This document is intended to outline our general product direction and is for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon. The development, release, and timing of any features or functionality described for Lhasa Limited’s products remains at the sole discretion of Lhasa Limited.