15 September 2018

Ministry of Defence

"Boxing Clever": Practical Techniques for Gaining Insights into Training Data and Monitoring Distribution Shift

Rob Ashmore, Matthew Hill

Defence Science and Technology Laboratory

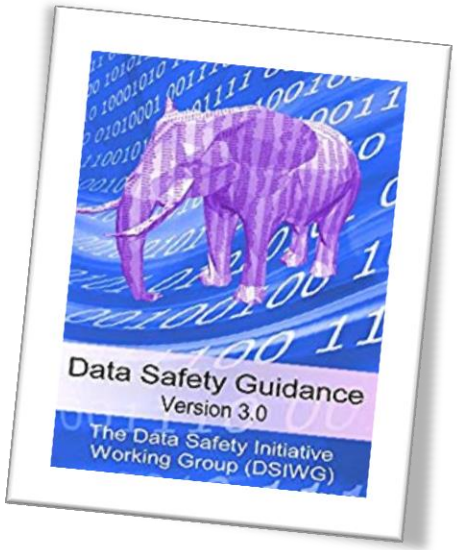
Disclaimer

This presentation is an overview of UK MOD sponsored research and is released for informational purposes only. The contents of this presentation should not be interpreted as representing the views of the UK MOD, nor should it be assumed that they reflect any current or future UK MOD policy. The information contained in this presentation cannot supersede any statutory or contractual requirements or liabilities and is offered without prejudice or commitment.

Key Concept

Training data is important; it should be analysed and justified as an item in its own right

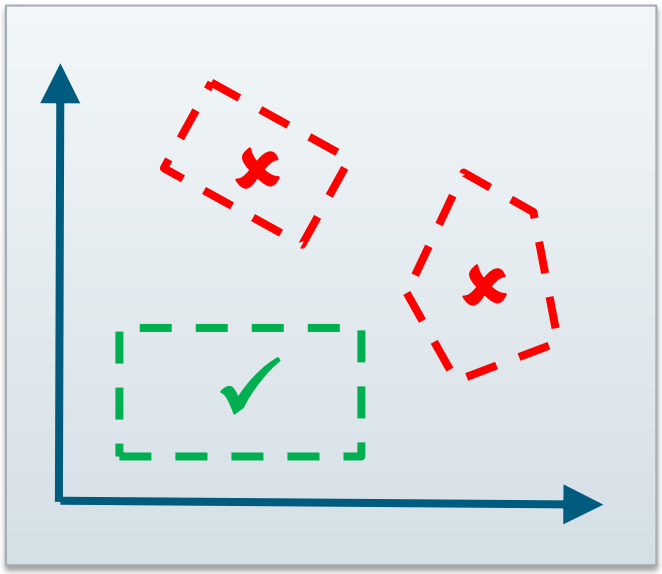
- Firstly, and most importantly, this data implicitly encodes the requirements for Machine Learning (ML) approaches
- Secondly, the importance of data, as an explicit item, is being recognised in "traditional" safety applications



Specific Motivating Example

- ~100k samples in data set, ~10 quantified features
  - Good surrogate data sets in the UCI ML repository
- Basically, our approaches focus on data sets that can naturally be represented in Comma-Separated Value (CSV) format
- Also, highly desirable to have approaches that (for this size of data set) can run on, for example, a commodity laptop
- Extension to other types of data set represents potential future work
  - Perhaps using wavelet decomposition for images

Axis-Aligned Boxes

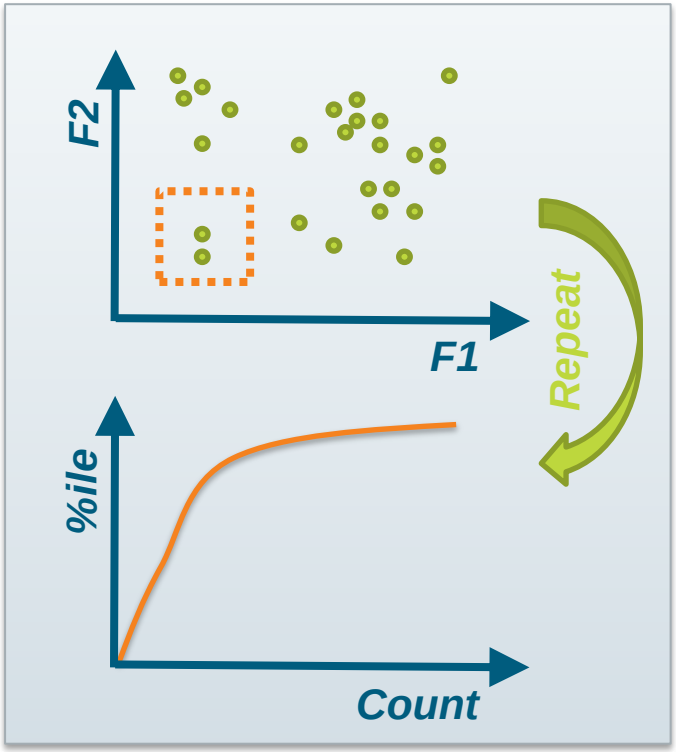


- Assume basic exploratory data analysis has been done
- We then make extensive use of axis-aligned boxes
- These implicitly define a way of measuring distance
  - Also imply some loose form of orthogonality between features
- Conscious trade between expressiveness and efficiency

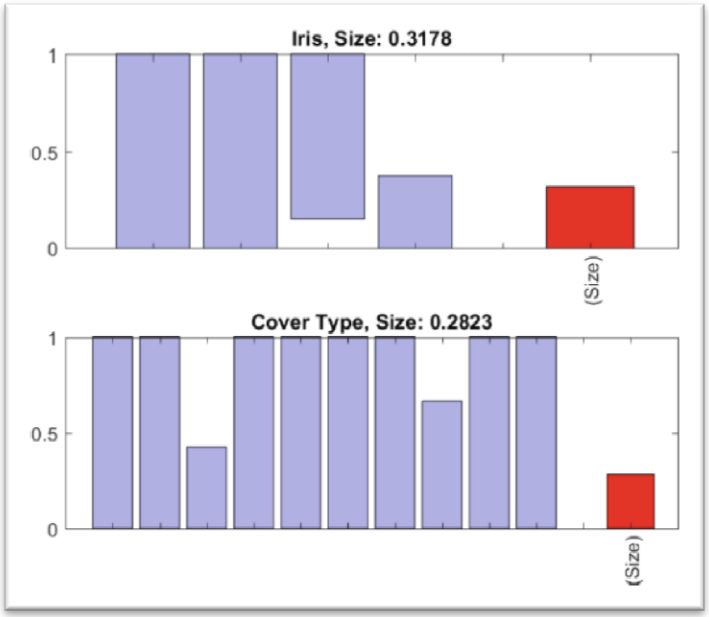
Axis-aligned boxes replace expensive distance calculations with cheap comparison operations

Sample Density

- Put an appropriately-sized, AA box around each sample; count number of samples in box; repeat
- Ratio between 1st and 99th percentiles of the distribution of counts is informative
  - Balance between densely and sparsely sampled regions
- Likewise, "lonely" and "popular" samples are also informative



Empty Hyper-Rectangles

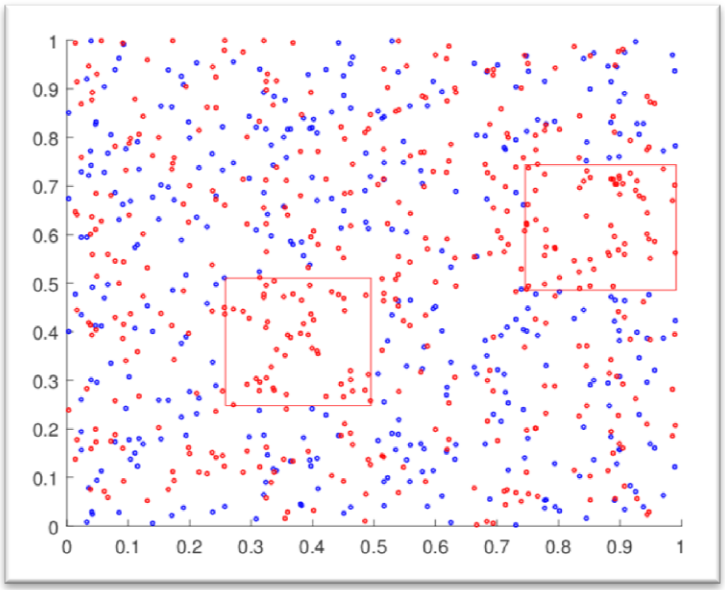


- Regions where there are no samples are interesting; these can be surprisingly large
- If training data is sampled from some external process, it can be difficult (or impossible) to fill these gaps

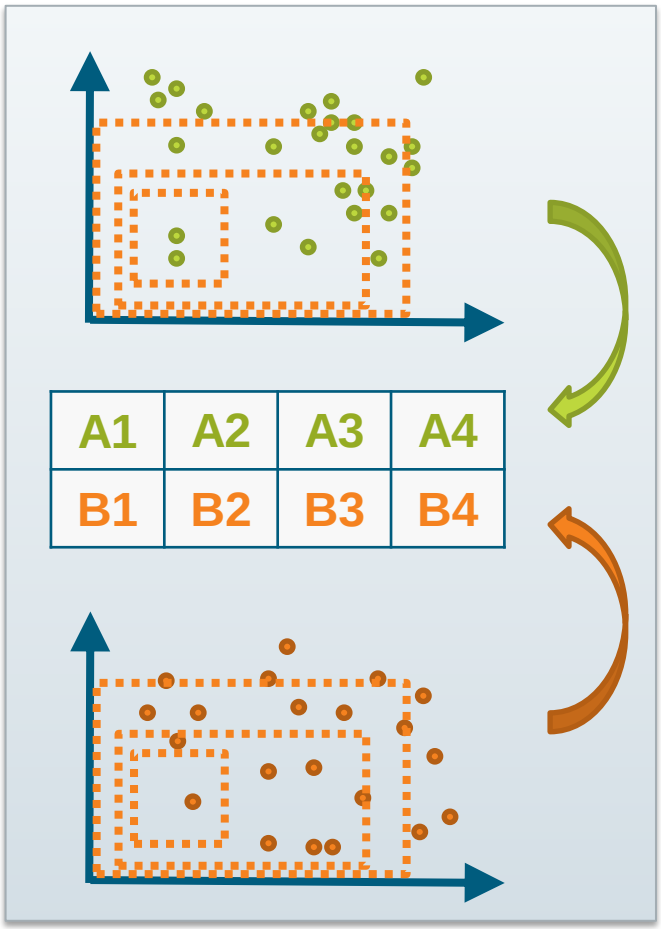
Knowing where there is no data is important

Single-Class Regions

- Regions that only contain samples from a single class are interesting
  - They provide a different way of testing the trained algorithm
  - Can be found using a simple extension to the EHR algorithm
- Single-class regions can cover very small fractions of the input domain, yet contain a significant proportion of samples in a class



Monitoring Distribution Shift



- For ***individual*** inputs: look at bounds of training data and EHRs
- For ***multiple*** inputs: create a nested set of AA boxes; count samples / inputs in each "layer"; run a χ^2 test
  - Simple to implement, fast to calculate
  - Can be used as a "canary" to trigger a more detailed calculation

Check operational inputs "match" training data

AA - Axis-Aligned
EHR - Empty Hyper-Rectangle

Summary

- Training data is important; it should be considered as an item in its own right
- Axis-aligned boxes allow for efficient computation, yet they can still be informative
- Sample density, EHRs, single-class regions and distribution shift can all be investigated
- Approaches have been used on real data: provided insight; informed an embryonic assurance case

15 September 2018

Ministry of Defence