An attempt to efficiently determine whether two data sets are “equivalent”

Instrument Analysis Workshop
February 28, 2006

GLAST LAT

Felix Schmitt (speaker)
Bijan Berenji
Elliott Bloom
Contents

• Motivation
• Cross Validation and classification trees; an estimator whether two data sets are equal
• Application to artificial data
• Application to MC GLAST data
• Outlook
Task: Get the photons, discard the background

Method 1 (successful): Classification Trees (B. Atwood[1])

- Train Classification Tree with MC: all_gamma and background
- Run real data through CT

Method 2 (also successful): Manually (E. Bissaldi[2])

- Results from Method 1
- Use MC to compare with real data
- Make cuts, using physical insight

Method 3 (not even close): Manually, enhanced (Berenji, Bloom, Schmitt)

- Like Method 2: make MC all_gamma and real data equal, making physically intuitive cuts
- Then: use Mechanism(?) to see if they differ and where


(?) Need suitable Mechanism to find differences between two datasets => this talk.
Motivation

Is MC data “equivalent” to GLAST ground data?

Ideal algorithm:
1. Two sets of bins: $n$ bins per variable
2. “Fill” bins with MC and GLAST ground data
3. Define measurement to compare bin filling topology

But: 269 variables (think MeritTuple) → $n^{269}$ bins

Classification trees

With traditional methods, comparing two large datasets is a daunting task.
Classification and regression trees

Example: B. Atwood's background rejection[3]

- sample \( s \) out of training data
- training data: MC of all gamma and background
- \( y(s) = \begin{cases} 1 & (s \in \text{all gamma}) \\ 0 & (s \in \text{background}) \end{cases} \)

Feeding a sample of real data through the tree yields prediction (0 or 1)


A classification tree makes predictions on one variable (“\( y \)”) from a new dataset. It is built from a training dataset for which \( y \) is known.
Is MC data “equivalent” to GLAST ground data?

Classification trees

Algorithm:
1. Two data sets MCdata, Grounddata
2. response variable y; sample s out of {MCdata, Grounddata}
   \[ y(s) = \begin{cases} 
   \text{TRUE} & (s \in \text{MCdata}) \\
   \text{FALSE} & (s \in \text{Grounddata}) 
   \end{cases} \]
3. generate CT from \( y \sim \{\text{MCdata, Grounddata}\} \)
4. Can CT distinguish between MCdata and Grounddata?

Point 4 is not yet clear: explanation follows

A C.T. is constructed and used to find differences between two datasets
Quality of classification trees[3]

Breiman et al.[4]:

- **complexity parameter** $cp$ (complexity punished growing/pruning):

  Abort tree growth when:
  
  \[
  \text{Classification error improvement} - cp \times \text{# of terminal nodes} < 0
  \]

- **10-fold cross-validation** of each $T(cp)$

- **best tree**: generated by the $cp$ with least cross validation error $CVE$

- **standard error** $SE = \sqrt{\frac{s^2}{N}}$, with $s^2 = \langle CVE^2 \rangle - \langle CVE \rangle^2$

\[\Longrightarrow CVE + SE < 0.5 \Rightarrow \text{The two datasets are different.}\]

\[\not\Longrightarrow CVE \pm SE \approx 0.5 \Rightarrow \text{The two datasets are (not necessarily) equivalent.}\]


*The classification error (from cross validation) is a measure for equivalence.*
Test the algorithm: Create two hypothetical data sets

Common properties of simMCdata, simGLASTdata

- 300 variables
- generated from uniform random distribution between [0, 1]

Differences of simMCdata and simGLASTdata

- simMCdata: 10k events
- simGLASTdata: 5k events
- distribution difference in first variable

simGLASTdata and simMCdata are purely hypothetical datasets to test the C.T. They have ABSOLUTELY NO physical meaning.
Two fake data sets

- distribution in simGLASTdata slanted by \( \text{atan}(0.5) \)

**Distribution difference of the two fake datasets**
Check I: compare two equivalent data sets

- simMCdata is randomly split in half and compared to itself

As expected, C.T.s are not able to find a difference between two equal data sets
Check II: does the C.T. find our prepared difference?

- simMCdata is compared to simGLASTdata

*The C.T. found a difference between the two fake datasets with different histograms.*
Check III: They are different, but where?

C.T.s also give (limited) information about where the differences originate.
Reality: compare (actual) MC data to itself

As expected, C.T.s are not able to find a difference between two equal data sets

- the first 100k events from all_gamma_10Mev20GeV_4M_merit
- split in half, compared to itself
Why I like R:

The entire code for everything I have said so far is exactly this:

```r
# this grows me the classification tree:
f <- rpart(y ~ data, method="class", minbucket=25, cp=1e-5)

# due to some (of course undocumented) funkiness in the module
# rpart, the cross-val error gets scaled with the resub. error
# of the (left split) of the root node. Reverse this:

# plot out x-val classification error in dependence of cp:
plotcp(f)
```

BUT: only “documentation” of rpart is the source code itself :-(

The classification error (from cross validation) is a measure for equivalence.
Problems & Outlook

Problems:

- R memory consumption high: 1.3GB for 100,000 samples
- rpart may not grow trees optimally
- No pre-prepared ground data available yet

Outlook:

- choose another CT implementation (maybe in c/c++)
- try gbm or rforest package for more accuracy? (if needed)
- compare actual MCS and Ground data

Thank you:

- Elliott Bloom
- Eduardo do Couto e Silva
- Bijan Berenji