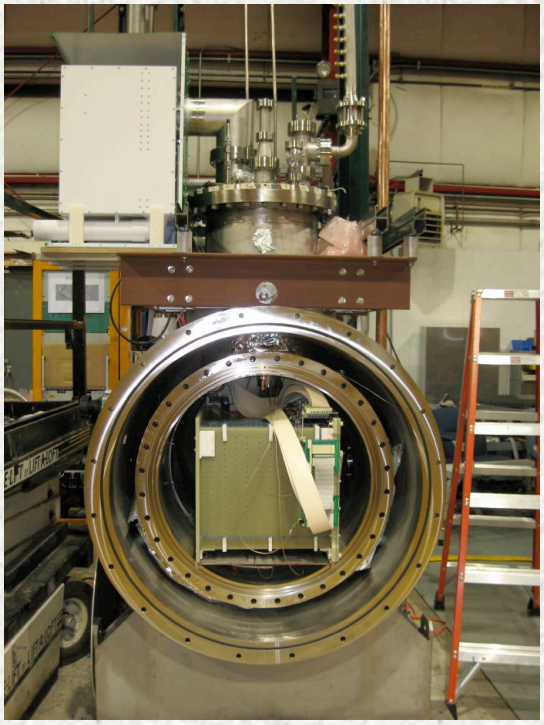
Multivariate Event Identification for a $\nu_\mu \rightarrow \nu_e$ Oscillation Experiment



Andrea Albert
Rice University

MicroBooNE Senior Thesis Presentation

May 8, 2009

Outline

- Intro:
  - Motivation
  - Detection Difficulties
  - Types of Events
- Visual Scanning Study
- Multivariable Event Identification Study

Introduction

Intro: Motivation

Why do we want to study $\nu_\mu \rightarrow \nu_e$ oscillations?

- determine mass hierarchy of neutrinos
- better measure mixing angle θ_{13}
- CP violation
  - could be studied by neutrino oscillations
  - understanding -> why matter/antimatter asymmetry

Intro: Detection Difficulties

- using $\sin^2(\theta_{23})=1/2$, optimal L and E_ν, and $\sin^2(2\theta_{13})=0.14\text{-}0.18$, $P(\nu_\mu \rightarrow \nu_e)=7\%\text{-}9\%$

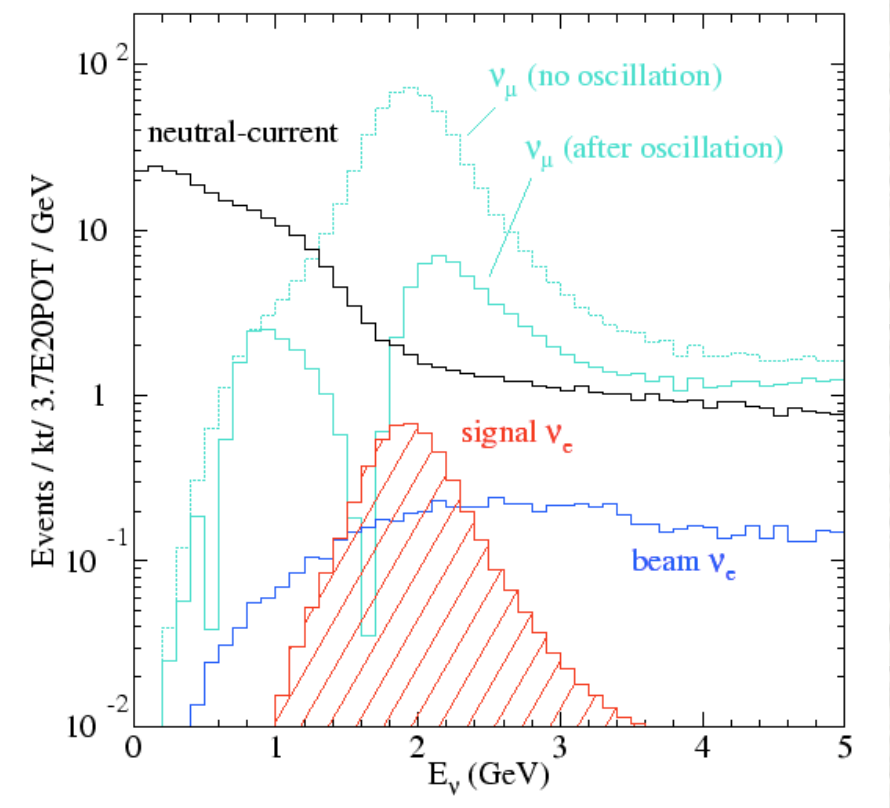
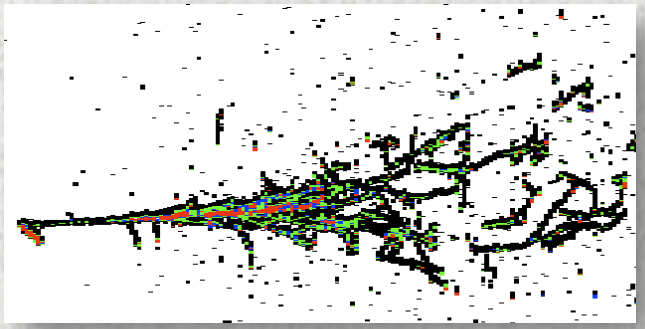


Fig 5.1- Simulated energy distributions for the oscillation signal, intrinsic beam events, neutral-current events and ν charged-current events with and without oscillations. The simulation used $\Delta m^2_{32} = 2.5*10^{-3}$ eV^2, $\sin^2(2\theta_{23})$, and $\sin^2(2\theta_{13})=0.04$. An off-axis distance of 12 km at 810 km was assumed. Source: NOνA proposal [11]

Intro: Types of Events

- Charged Current ν_e (CCe)

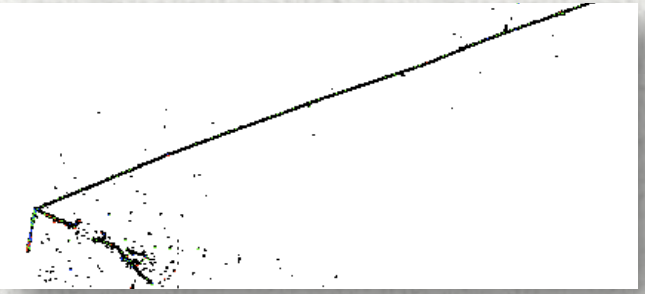
  $\nu_e + n \rightarrow e^- + p$



CCe Quasi-Elastic Scattering

- Charged Current ν_μ (CCmu)

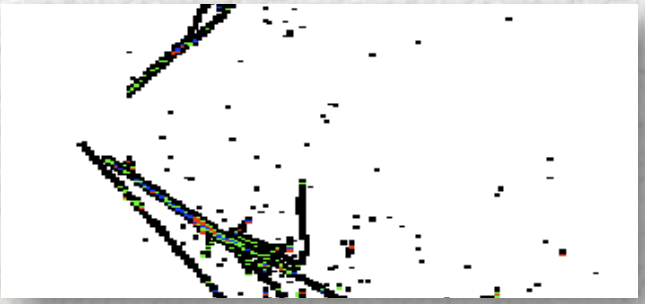
  $\nu_\mu + n \rightarrow \mu^- + p$



CCmu Resonance Scattering

- Neutral Current (NC)

  $\nu + N \rightarrow \nu + N$



NC Deep Inelastic Scattering

Images from Monte Carlo simulations

Visual Scanning Study

Visual Scanning: Procedure

- visually classified CCe, CCmu, or NC events from MC simulations using different wire spacings
- Goal: to determine if efficiency is significantly increased with closer wire spacing
- $E_{events} < 3.5GeV$
- Only scanned events where e^- materilized within 2 cm and 5 cm of PV

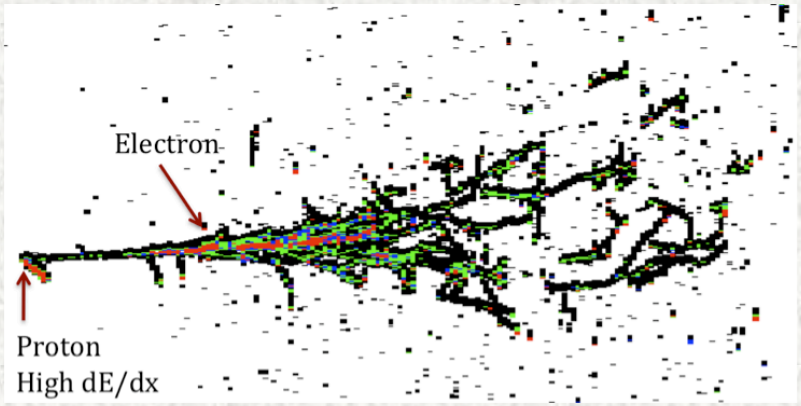
Visual Scanning: General Method

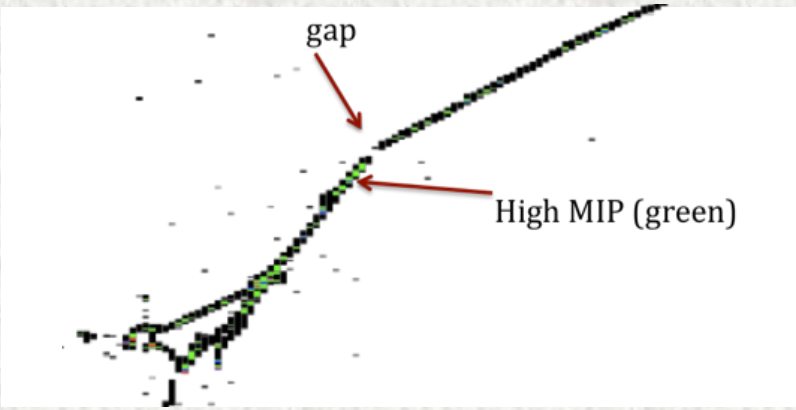
- CCe ID: look for e^- produced at PV

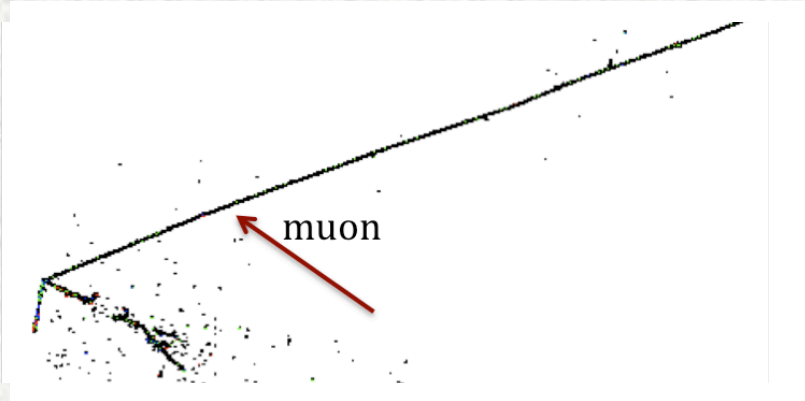
- e^- ID:
  - EM shower
  - MIP=1 at beginning
- μ^- ID:
  - long straight track
  - MIP=1 mostly
  - possible decay into e^-

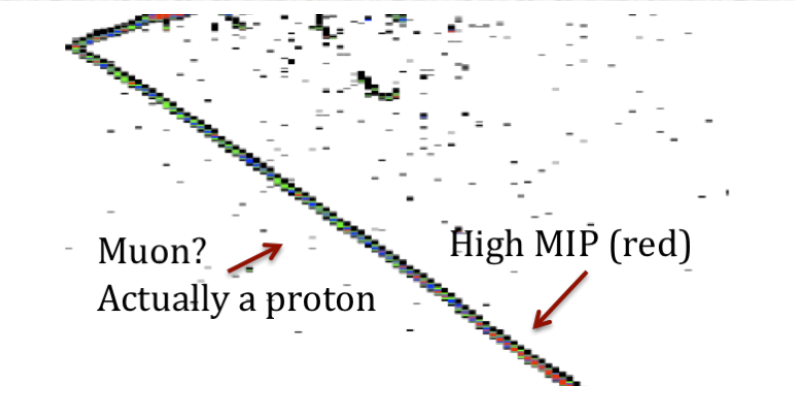
- p^+ ID:
  - High dE/dx
- γ conversion ID:
  - gap btw track beginning and PV
  - MIP=2 at beginning

Visual Scanning: General Method









Visual Scanning: Results

e^- materialized 2 cm from PV

5mm pitch

train05_1.eye
Num: CCe 111 CCmu 379 CCtau 0 NC 1509
Number of events 93

*** Your selection ***

	CCe	CCmu	NC
true CCe	64	0	3
true CCmu	1	3	0
true NC	7	2	13

CCe Efficiency **95.5%** NC rejection **99.54%** CCmu rejection **99.74%**

10mm pitch

train10_1.eye
Num: CCe 115 CCmu 399 CCtau 0 NC 1485
Number of events 103

*** Your selection ***

	CCe	CCmu	NC
true CCe	75	0	1
true CCmu	0	2	0
true NC	11	2	12

CCe Efficiency **98.7%** NC rejection **99.26%** CCmu rejection **100.00%**

Visual Scanning: Results

e^- materialized 5 cm from PV

5mm pitch

trainv2_05_2.eye
Num: CCe 126 CCmu 384 CCtau 0 NC 1489
Number of events 154

*** Your selection ***

	CCe	CCmu	NC
true CCe	86	2	6
true CCmu	0	5	4
true NC	6	3	42

CCe Efficiency **91.5%** NC rejection **99.60%** CCmu rejection **100.00%**

10mm pitch

trainv2_10_2.eye
Num: CCe 157 CCmu 399 CCtau 0 NC 1443
Number of events 176

*** Your selection ***

	CCe	CCmu	NC
true CCe	80	5	17
true CCmu	0	12	1
true NC	11	3	47

CCe Efficiency **78.4%** NC rejection **99.24%** CCmu rejection **100.00%**

Visual Scanning: Conclusion

- closer spacing slightly increased bkgd reduction (< 0.5%)
- when scanned e^- produced 5 cm from PV, CCe ID efficiency significantly increased with 5mm spacing (78.4% → 91.5%)

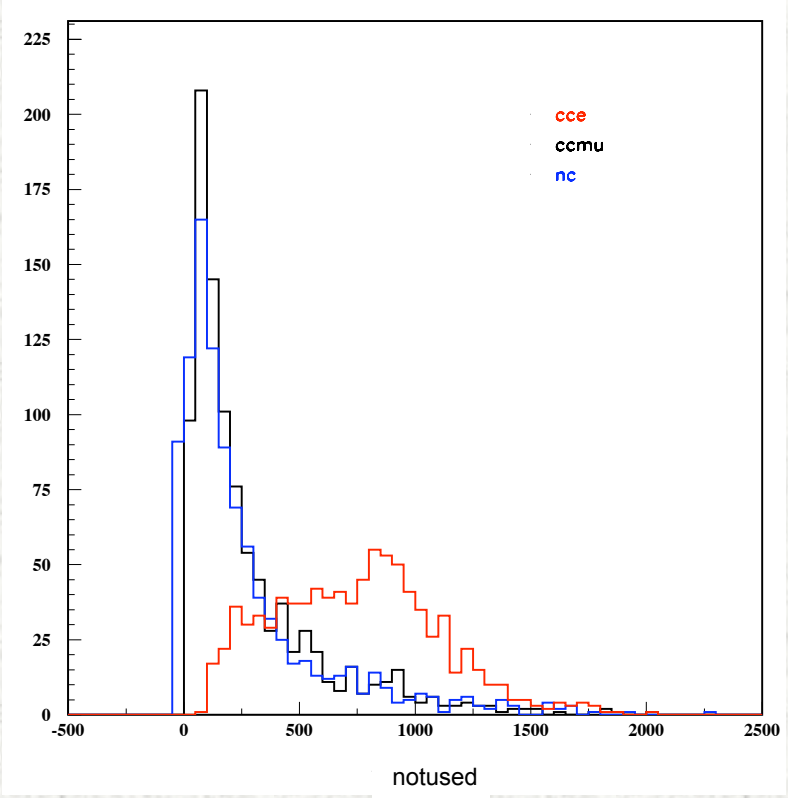
Multivariate Study

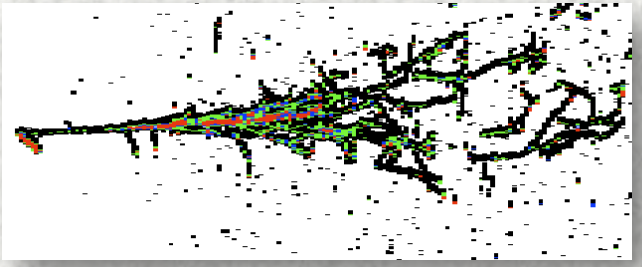
Multivariate Study: Procedure

- Use Monte Carlo simulations to establish distinguishing variables
- Note: ArgoNeuT simulation used produces two 2D views of event
- Studied events with visible energy < 4GeV
- Made optimized cuts in PAW
- Trained a C4.5 Decision Tree

Multivariate Study: My Variables

- ncluster: number of clusters*
- nhit: number of hits
- notused: number of hits not fitted to cluster
- tmip: sum of minimum ionizing particle values
- long: number of hits in longest cluster



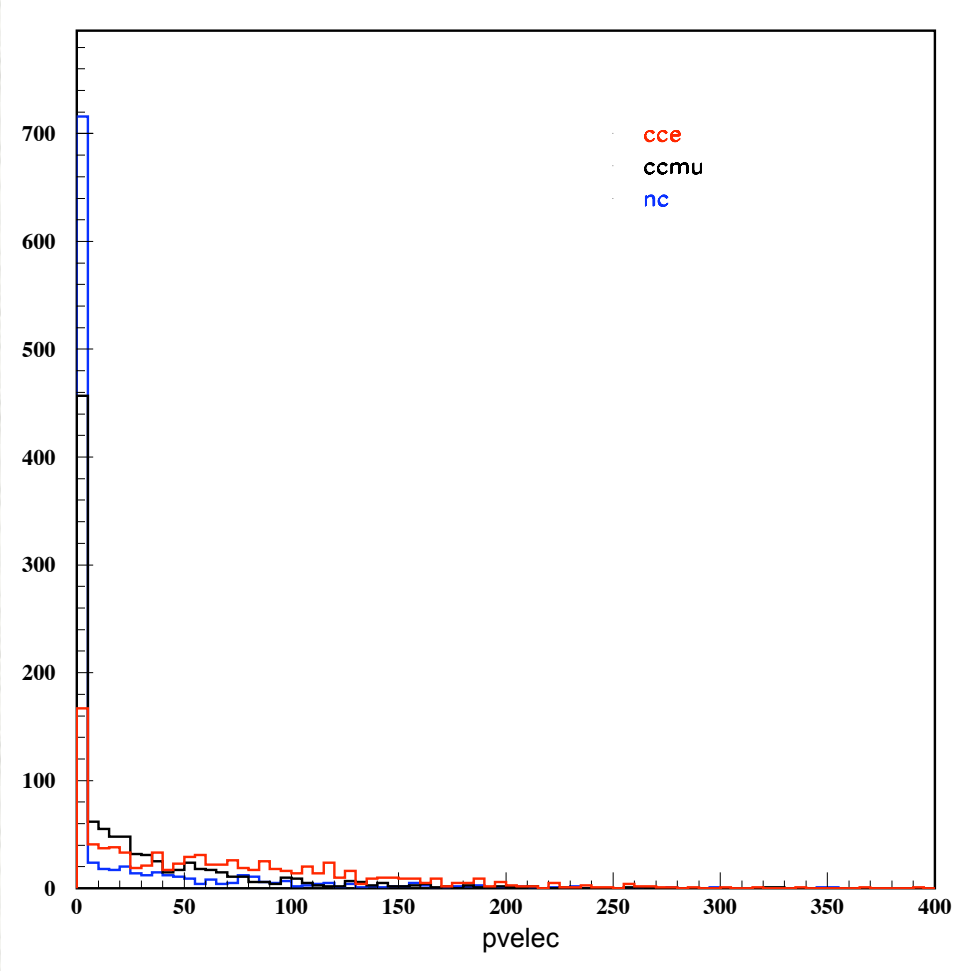


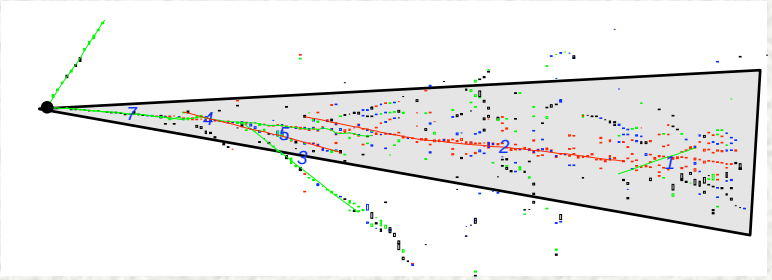
sample signal event

*A cluster is a line-like collection of hits in one detector view

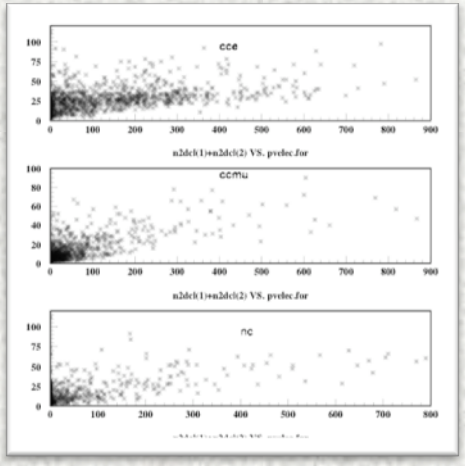
Multivariate Study: My Variables

- Number of hits in a triangle with opening angle 10^0 centered on each primary vertex cluster
- > 0 if primary vertex e^-

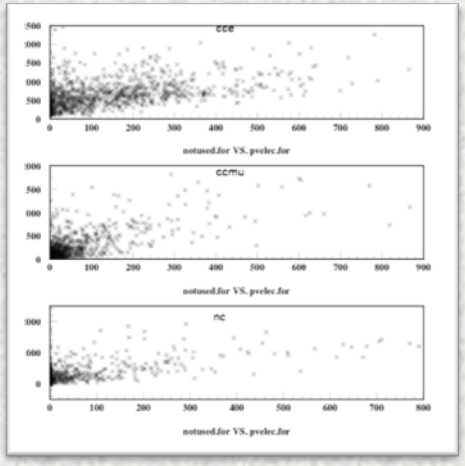




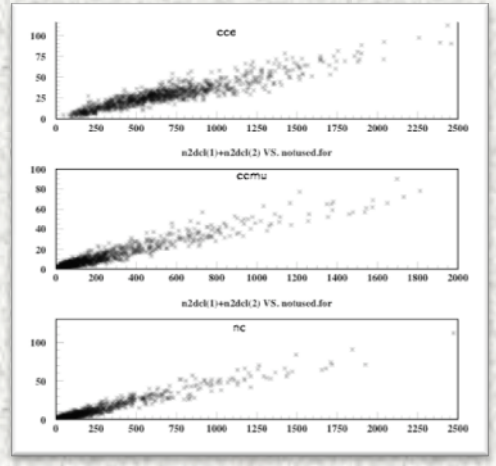
Variable Correlations



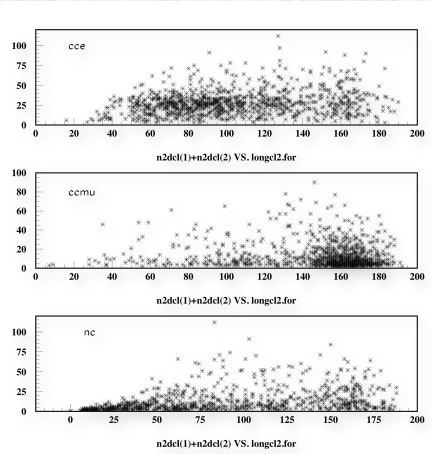
ncluster vs pvelec



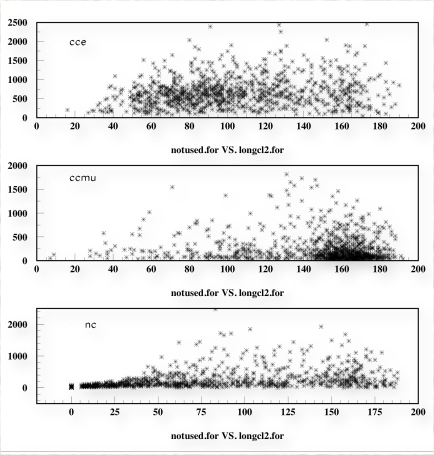
notused vs pvelec



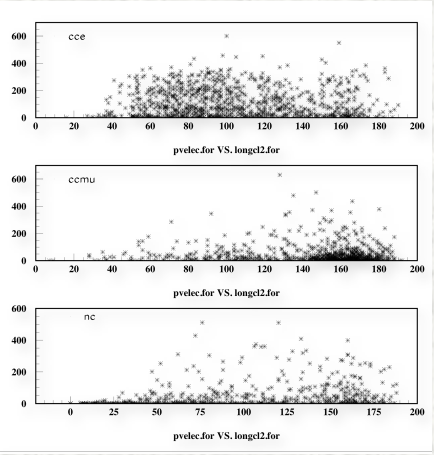
ncluster vs notused



ncluster vs long



notused vs long



pvelec vs long

Optimized Cuts in PAW

- Used ncluster, notused, pvelec, and long because mostly uncorrelated and utilized every hit
- Found which cuts maximize $\frac{signal}{\sqrt{signal + bkgd}}$
- ncluster > 2, notused > 235, pvelec > 0, long < 182

avg CCe efficiency	75.04%
avg CCmu misidentification	29.05%
avg NC misidentification	17.49%
avg bkgd misidentification	23.27%

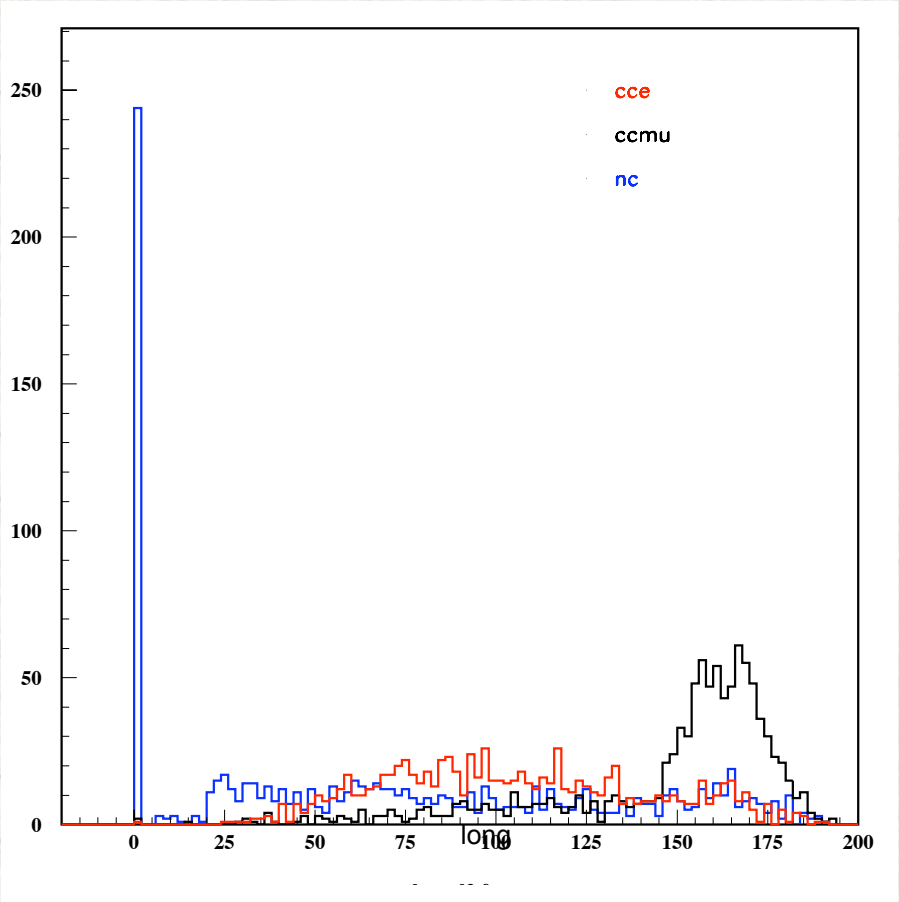
Results from optimized cuts: ncluster > 2, notused > 235, pvelec > 0, long < 192

Utilizing Long

- A lot of CCmu background got through
- changed long cut to be 25 < long < 145

avg CCe efficiency	60.17%
avg CCmu misidentification	7.76%
avg NC misidentification	8.37%
avg bkgd misidentification	8.07%

Results from cuts: ncluster > 2, notused > 235, pvelec > 0, 25 < long < 145



Optimized Cuts with 7% CCe Data Set

- CCe signal will be ~7% of total events
- Data sets contain 1000 CCmu, 550 NC, and 117 CCe
- Background still dominates over signal

Data Set 1

	CCe	Bkgd	←classified as
CCe	**70**	47	
CCmu	**84**	916	
NC	**58**	492	
↑actual			

Data Set 2

	CCe	Bkgd	←classified as
CCe	**68**	49	
CCmu	**98**	902	
NC	**53**	497	
↑actual			

Data Set 3

	CCe	Bkgd	←classified as
CCe	**64**	53	
CCmu	**96**	904	
NC	**63**	487	
↑actual			

Data Set 4

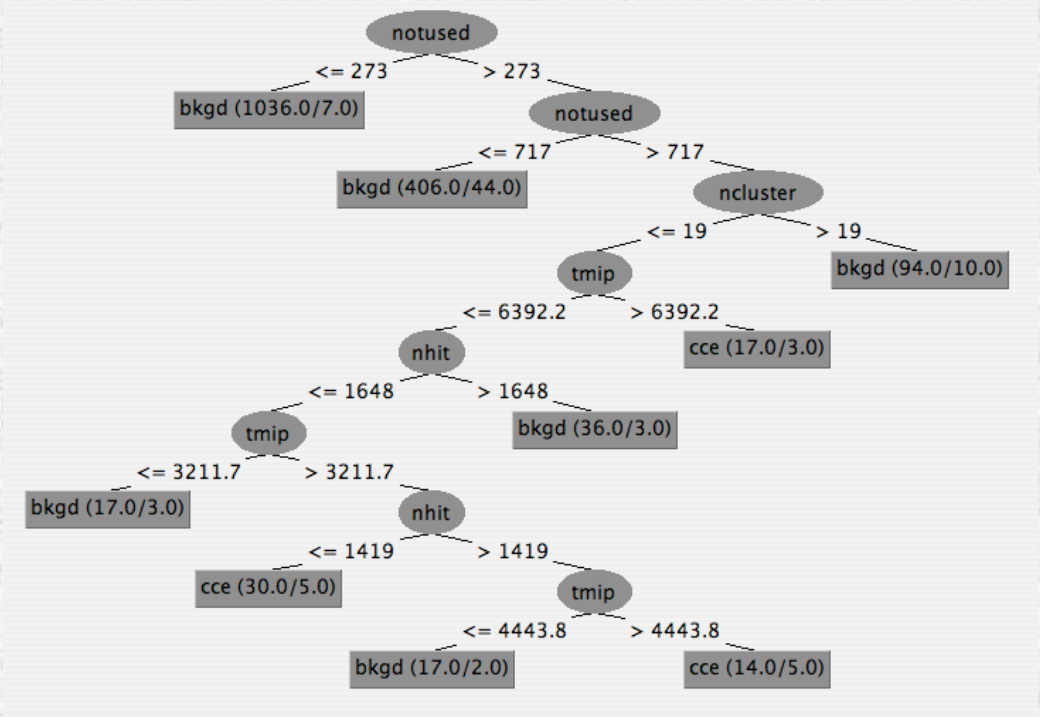
	CCe	Bkgd	←classified as
CCe	**67**	50	
CCmu	**85**	915	
NC	**54**	498	
↑actual			

avg CCe efficiency	57.48%
avg CCmu misidentification	9.08%
avg NC misidentification	10.35%
avg bkgd misidentification	9.72%

Results from optimized cuts using 7% CCe data sets: ncluster > 2, notused > 235, pvelec > 0, long < 192

Decision Trees

- Training -> sophisticated cuts, which should classify test sets well
- Used a C4.5 tree, which maximizes information gained per node $Info([N_1, N_2, \ldots, N_k]) = -\sum_k \frac{N_k}{S} \log_2 \frac{N_k}{S}$



C4.5 tree trained with 7% CCe data

Initial Decision Tree Test

- Most successful using correlated variables ncluster, notused, nhit, and tmip
- CCmu and NC combined as bkgd
- Trained on 66% of data set and tested on remaining 34%
- High signal purity, ok background reduction

avg CCe efficiency	76.66%
avg bkgd misidentification	17.25%

Decision tree results from data sets with equal numbers of CCe, CCmu, and NC events

Decision Tree with 7% CCe Data Set

- Trained on 1 set with 7% CCe signal and tested on remaining 3 sets
- Results in avg signal:false-signal of 1.72:1
- Final events classified as signal are mostly actual signal

Data Set 2

	CCe	**Bkgd**	←classified as
CCe	**32**	85	
Bkgd	**16**	1534	
↑acutal			

Data Set 4

	CCe	**Bkgd**	←classified as
CCe	**30**	87	
Bkgd	**19**	1531	
↑acutal			

Data Set 3

	CCe	**Bkgd**	←classified as
CCe	**33**	84	
Bkgd	**21**	1529	
↑acutal			

avg CCe efficiency	27.07%
avg bkgd contamination	1.20%

Decision tree results from train on 1 7% CCe data set and training on the other 3

Conclusion

- Visually scanning events with an e^- 5 cm from PV with 5mm spacing instead of 10mm
  - greatly increased CCe ID efficiency (78.4% → 91.5%)
  - slightly increased NC rejection (99.24%→99.60%)
- Training a C4.5 Decision Tree with the 7% CCe data resulted in a final signal:false-signal of 1.72:1 with the remaining CCe data (most events labeled as CCe actually CCe!)
- More tests need assuming lower $P(\nu_\mu \rightarrow \nu_e)$
- More variables focusing on PV may be necessary

A big thanks to…

- Marj Corcoran
- Bruce Baller
- Stephen Pordes and Hans Jöstlein
- Zigfried Hampel-Arias
- Rice University Physics and Astronomy Department

Bibliography

- C. Anderson *et al.* [MicroBooNE collaboration] "A Proposal for a New Experiment Using the Booster and NuMI Neutrino Beamlines: MicroBooNE" October 2, 2007.
- The NOvA Collaboration "NuMI Off-Axis ν_e Appearance Experiment" March 21, 2005
- R. Rameika. "LAr5- A Liquid Argon Neutrino Detector for Long Baseline Neutrino Physics" March 27, 2008 PAC Presentation
- ArgoNeuT: Mini LArTPC Exposure to Fermilab's NuMI Beam <http://t962.fnal.gov/index.html>
- The Particle Data Group, C. Amsler *et al*. Phys. Lett. **B667** (2008)
- "Neutrino Oscillation". Wikipedia <http://en.wikipedia.org/wiki/Neutrino_oscillation>
- E. Frank and I. H. Witten. Data Mining. Morgan Kaufmann Publishers, 2000

  <http://www.cs.waikato.ac.nz/ml/weka/>

Backup Weka Slides

Weka

- Java program developed by The University of Waikato
- Allows one to test with different machine learning algorithms (different decision trees, neutral nets, etc) within the same program

Weka: How to

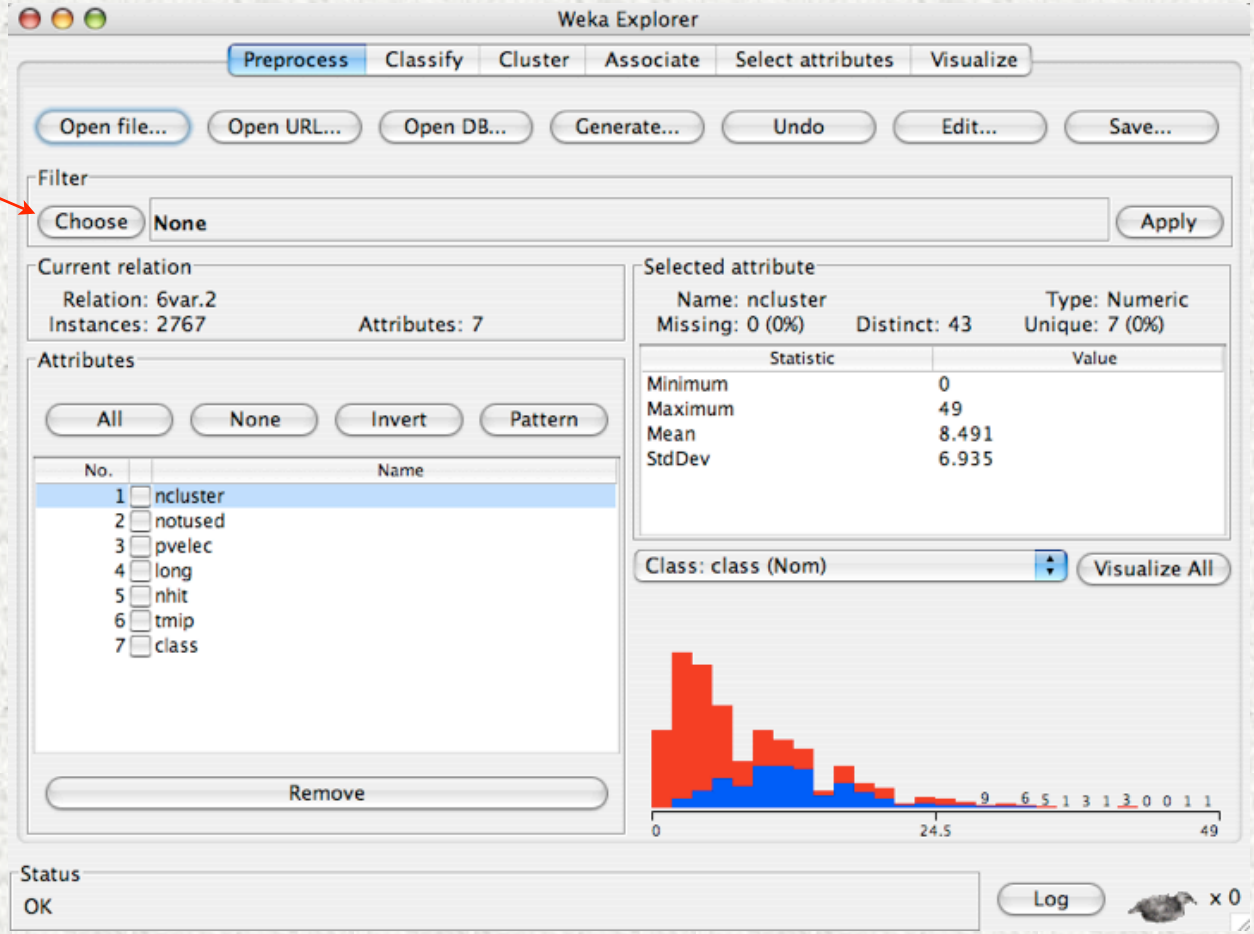
- requires data to be in .arff format
- attribute = distinguishing variable
- class = what events can be classified as
- instance = individual event

6var_0.07cce.arff

```
@relation 6var_0.07cce

@attribute ncluster numeric
@attribute notused numeric
@attribute pvelec numeric
@attribute long numeric
@attribute nhit numeric
@attribute tmip numeric
@attribute class {cce,bkgd}

@DATA

 3.0, 116.0,  13.0,147.0, 448.0, 835.7,bkgd
 2.0,  12.0,   0.0,157.0, 309.0, 465.2,bkgd
 7.0, 162.0,  81.0,172.0, 662.0,1489.7,bkgd
 6.0, 265.0,  51.0,164.0, 720.0,1690.9,bkgd
 2.0,  44.0,   0.0,127.0, 293.0, 536.9,bkgd
 8.0, 114.0,   0.0,172.0, 855.0,1380.7,bkgd
10.0, 542.0,  78.0,150.0,1215.0,2452.1,bkgd
 3.0,  68.0,   0.0,155.0, 392.0, 636.1,bkgd
```

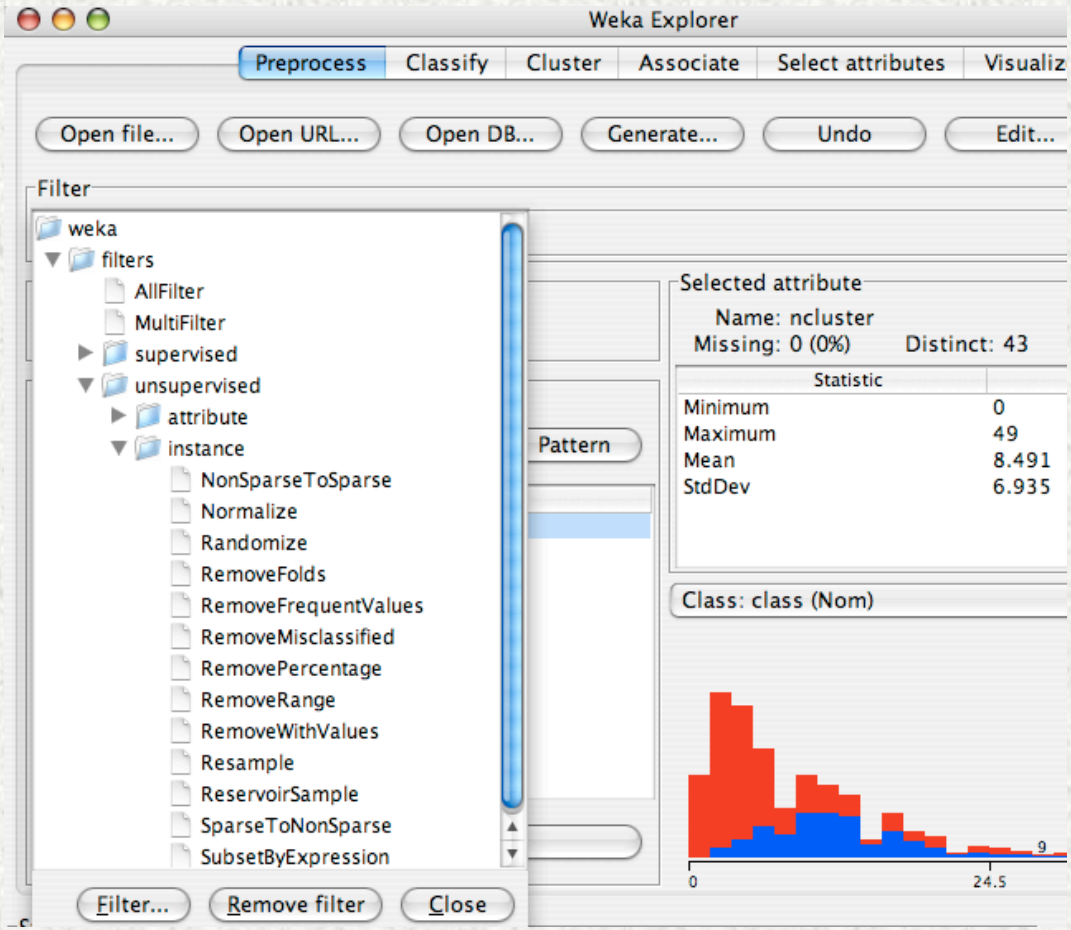
Weka: How to

choose filters



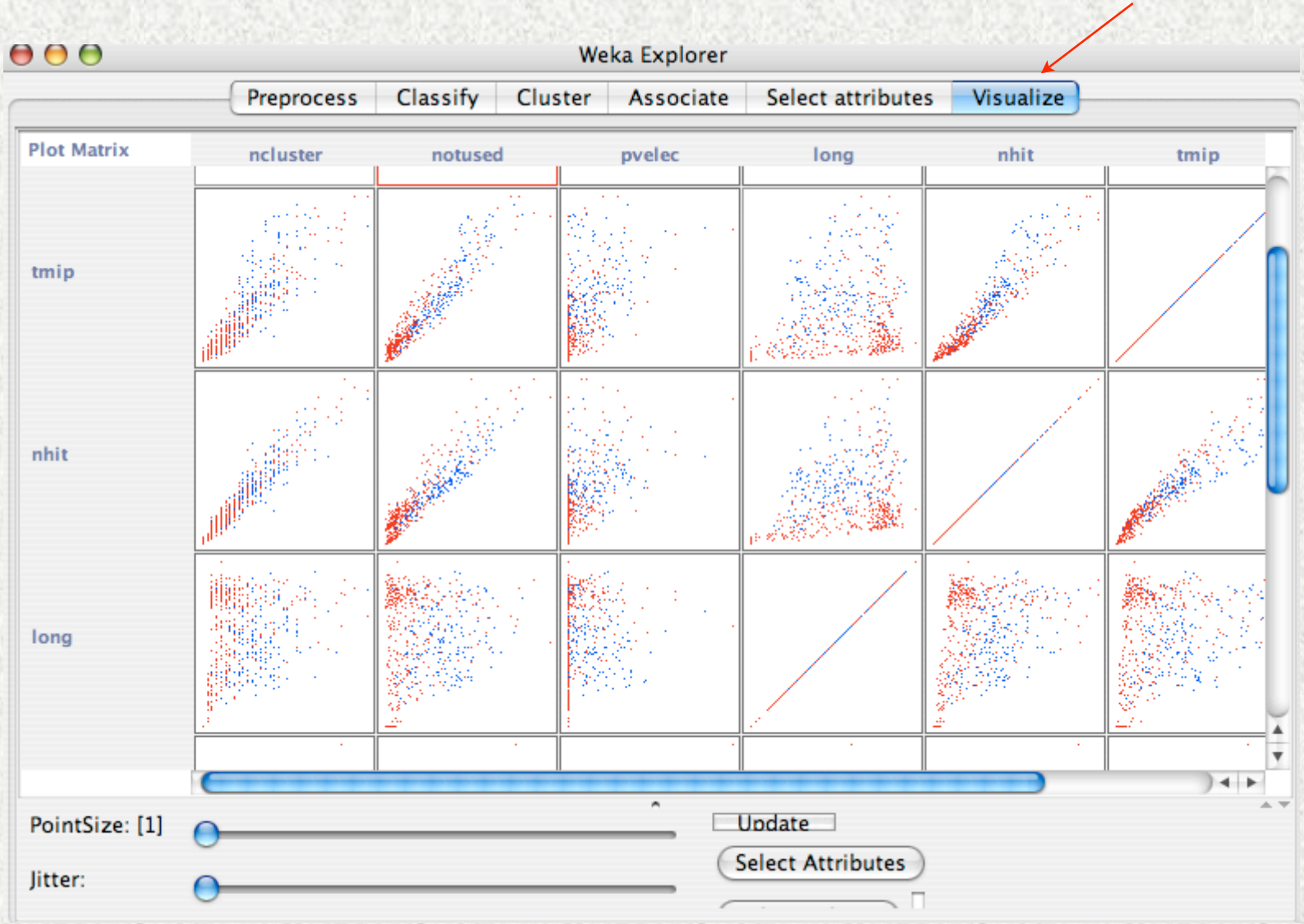
Main Explorer Window

Weka: How to



Can choose different filters

Weka: How to


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
ncluster
notused
pvelec
long
nhit
tmip
tmip
nhit
long
PointSize: [1]
Jitter:
Update
Select Attributes

Weka: How to

Choose learning algorithm

specify training and testing data sets

change algorithm variables

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose J48 -C 0.25 -M 10

Test options

- Use training set
- Supplied test set — Set...
- Cross-validation — Folds 10
- Percentage split — % 66

More options...

(Nom) class

Start | Stop

Result list (right-click for options)

10:30:30 - trees.J48

Classifier output

```
Size of the tree :      45

Time taken to build model: 0.11 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances          785               83.4219 %
Incorrectly Classified Instances        156               16.5781 %
Kappa statistic                           0.6108
Mean absolute error                       0.2518
Root mean squared error                   0.3655
Relative absolute error                  57.9136 %
Root relative squared error              78.3711 %
Total Number of Instances               941

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure
                   0.701     0.103       0.762    0.701       0.73
                   0.897     0.299       0.864    0.897       0.88
Weighted Avg.      0.834     0.236       0.832    0.834      0.832

=== Confusion Matrix ===

   a   b   <-- classified as
 211  90 |   a = cce
  66 574 |   b = bkgd
```

Classifying Window