



# Higgs to $ZZ^*$ Lepton State Decay on CEPC Based on Machine Learning Research

ZHANG RUIHAN



# Graduation Dissertation: Outline and Plans

Results run by LHC:

	Untagged	VBF	VH	ttH
H→γγ	✓	✓	✓	✓
H→ZZ→4l	✓	✓	✓	✓
H→WW→2l2ν	✓	✓	✓	✓
H→ττ	✓	✓	✓	✓
H→bb			✓	✓
H→μμ	✓	✓		

Decay mode	Branching fraction [%]
$H \rightarrow bb$	$57.5 \pm 1.9$
$H \rightarrow WW$	$21.6 \pm 0.9$
$H \rightarrow gg$	$8.56 \pm 0.86$
$H \rightarrow \tau\tau$	$6.30 \pm 0.36$
$H \rightarrow cc$	$2.90 \pm 0.35$
$H \rightarrow ZZ$	$2.67 \pm 0.11$
$H \rightarrow \gamma\gamma$	$0.228 \pm 0.011$
$H \rightarrow Z\gamma$	$0.155 \pm 0.014$
$H \rightarrow \mu\mu$	$0.022 \pm 0.001$

Run by CEPC:

- Higgs can be detected by recoil mass → distinguish Higgs and its decay particles independent of any particular model. (Only Z boson is reconstructed)
- Cleaner tracks → A better measurement → Improved accuracy measurement of H-Z coupling, 10 times better than LHC.
- More Physical phenomena could be explored → Better measurement of invisible decay tracks → Higgs singular decay.



## What to achieve:

1. Filter collision cases in CEPC framework
2. Discriminate between signals and bkgd
3. Fit in distribution

## How to discriminate:

- Analyses momentum distribution:
- According to their different kinematics features, signal particles and bkgd have different momentum distribution
- Pinpoint the generation peak and use energy cut
- Machine Learning method



# Graduation Dissertation: Outline and Plans

## Decision Tree:

- Set of information: D
- Variables (Excepted classifications of D): C(1) C(2) C(3)···C(k)
- Property: A
- Variable (Classifications divided by a): D(1) D(2) D(3)···D(v)

- Entropy of D:

$$H(D) = - \sum_{k=1}^K \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|}$$

- Entropy of D based on property A:

$$H(D|A) = \sum_{i=1}^n \frac{|D_i|}{|D|} H(D_i) = - \sum_{i=1}^n \frac{|D_i|}{|D|} \sum_{k=1}^K \frac{|D_{ik}|}{|D_i|} \log_2 \frac{|D_{ik}|}{|D_i|}$$

← Expectation of D's entropy distribution to A, based on the precondition of A

- Gini rate:

$$Gini(D) = \sum_{k=1}^{| \gamma |} \sum_{k^1 \neq k} p_k p_{k^1} = 1 - \sum_{k=1}^{| \gamma |} p_k^2$$

← Randomly choose two samples from set D, the possibility that they are from two separated classification.



# Graduation Dissertation: Outline and Plans

- Gini index:

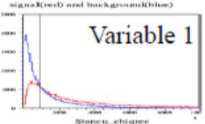
$$Gini_{index}(D,a) = \sum_{v=1}^V \frac{|D^v|}{|D|} Gini(D^v)$$

- Smaller Gini index means larger information purity
- The Toolkit for Multivariate Data Analysis with ROOT (*TMVA*)

## A Decision Tree

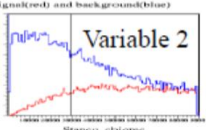
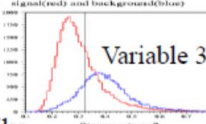
(sequential series of cuts based on MC study)

( $N_{signal}/N_{bkgd}$ )  
40000/40000



bkgd-like  
9755/23695

signal-like  
30,245/16,305



bkgd-like  
1906/16828

signal-like  
7849/6867

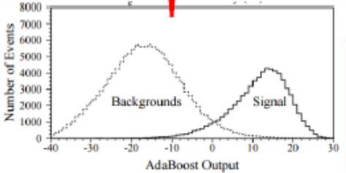
signal-like  
20455/3417

bkgd-like  
9790/12888

通过Boosting 算法不断提高误判事例的权重，产生一系列Decision Trees



把每个事例在所有Decision Trees获得的积分累加，通过“Majority vote”方法提高性能和稳定性。



通过Boosting不断提高误判事例的权重，使得这些难以区分的事例在后续的Decision Trees获得的正确区分，提高效率。