Ensemble of maximized Weighted AUC models for the maximization of the median discovery significance

Roberto Díaz-Morales - rdiazm@tsc.uc3m.es

University Carlos III de Madrid
Department of Signal Theory and Communications

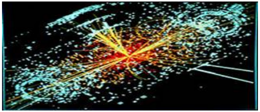
December, 2014

Table of Contents

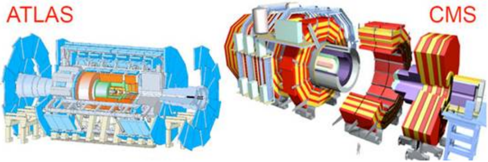
1. The challenge
   - The ATLAS and CMS experiments
   - The HEPML Challenge
   - The Dataset
   - The Goal
2. Preprocessing
   - Reducing the Feature Space
   - Missing Values Imputation
3. Training Procedure
   - AMS Maximization
   - Bootstrap Aggregating
   - Ensemble
   - Schema
4. Results
5. Conclusions
6. Acknowledgments

The ATLAS and CMS experiments

- At CERN, inside the LHC, proton beams are accelerated in both directions. Some protons collide producing a variety of different particles with a broad range of energies.



- ATLAS and CMS are two of the general purpose particle physics detectors built on the LHC.



- The ATLAS and CMS experiments claimed the discovery of the Higgs Boson through decay mechanisms.
- The dataset for this competition was provided by the official ATLAS detector simulator.

The HEPML Challenge

The Higgs Boson Machine Learning Challenge was organized to encourage the collaboration between high energy physicist and data scientist.

The Challenge was was hosted by Kaggle

It took place from May 12th 2014 to September 15th 2014.

It brought together 1785 teams.

The Dataset

$$\mathcal{D} = \{(\mathbf{x}_1, y_1, w_1), ..., (\mathbf{x}_n, y_n, w_n)\}$$

$$y_i \in \{background, signal\}$$

$$\mathbf{x}_i \in \mathbb{R}^d$$

$$w_i \in \mathbb{R}^+$$

- Events divided into two categories.
- A weight associated to every event.
- Dataset size n=250000
- Number of features d=30. (17 primitive and 13 derived)
- Missing values.

Evaluation metric

The goal was, given the feature space, to find the region where the signal events are signifiantly higher than the background events.

The evaluation metric was the approximate median significance (AMS):

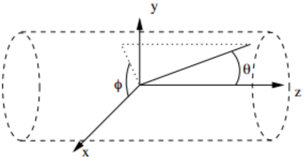
$$AMS = \sqrt{2\left((s+b+b_r)log\left(1+\frac{s}{b+b_r}\right)-s\right)} \approx \frac{s}{\sqrt{b}}$$

$$s = \sum_{i=1}^{n} w_i 1\{y_i = s\}1\{\hat{y}_i = s\}$$

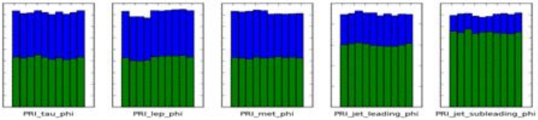
$$b = \sum_{i=1}^{n} w_i 1\{y_i = b\}1\{\hat{y}_i = s\}$$

$$b_r = 10$$

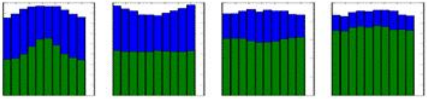
Reducing the Feature Space



Rotation and change of direction to make the events invariant to the angle ϕ and direction of the z axis. Reduces by 2 the number of features.
For example, the 5 ϕ angles:



Become:



Missing values

Some events have missig values, for example:

- Jets can appear 0, 1, 2 or 3 times in an event, some features are undefined depending the number of jets:
    - The azimuth angle of a jet is undefined if the number of jets is 0.
    - The pseudorapidity separation between jets is undefined if the number of jets is 0 or 1.
    - ...
- The estimated mass of the Higgs boson candidate may be undefined.

There are eight different kinds of patterns if we look at the number of jets and the estimated mass.

Every missing data has been replaced by his mean and eight binary features has been added to indicate what kind of pattern is.

AMS Maximization

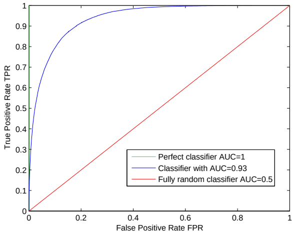
Maximizing the AMS is not an easy task:

- It is not differentiable.
- Non convex.

Can be approximated as:

$$AMS \approx \frac{s}{\sqrt{b}}$$

Let's try a different approach, finding the parameters that obtain the highest AUC.



$$WTPR = \frac{\sum_{i=1}^{n} w_i\{y_i = s\}\{\hat{y}_i = s\}}{\sum_{i=1}^{n} w_i\{y_i = s\}} = \frac{s}{max(s)}$$

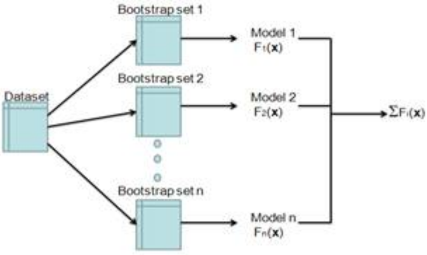
$$WFPR = \frac{\sum_{i=1}^{n} w_i\{y_i = b\}\{\hat{y}_i = s\}}{\sum_{i=1}^{n} w_i\{y_i = b\}} = \frac{b}{max(b)}$$

Bootstrap Aggregating

Working with the dataset I observed some instability in the results:

- Small changes in the parameters → Very different results
- Removing a small number of samples → Very different results

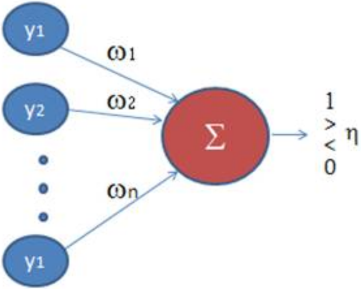
Bootstrap aggregating, also calling bagging, is a technique to improve the stability of machine learning algorithms.



It reduces the variance of the results as well as help to avoid overfitting.

Ensemble

Ensemble methods use the output of multiple learningclassifiers to obtain a better predictive performance than a single classifier.
Empirically, these techniques are able to achieve better results if there is diversity among the classifiers.



To avoid overfitting problems I chose a linear combination of model because of its simplicity. The weights $w = \{w_1, .., w_l\}$ and the threshold η have been chosen to maximize the AMS.

The Algorithm

Given a data set with n samples $\mathcal{D} = \{(\mathbf{x}_1, y_1, w_1), ..., (\mathbf{x}_n, y_n, w_n)\}$

Step 0: Initialization

- A set of m training sets is generated using different transformations. $\mathbf{D} = \{\mathcal{D}_1, ..., \mathcal{D}_m\}$, where $\mathcal{D}_i = \{(f_i(\mathbf{x}_1), y_1, w_1), ..., (f_i(\mathbf{x}_n), y_n, w_n)\}$
- d possible algorithms.
- The number of possible models is then $s = md$.

Step 1: Parameter selection

The parameters every model have been chosen to achieve the maximum WAUC.

Step 2: Bagging

Every model has been trained using bagging to stabilize results.

Step 3: Ensemble

The final classifier is built as a linear combination of l the models:

$$f(\mathbf{x_i}) = \sum_{i=1}^{l} w_i f_i(\mathbf{x_i}) \overset{1}{\underset{0}{\gtrless}} \eta$$

$\mathbf{w} = \{w_1, ..., w_l\}$ and η are obtained to maximize the AMS.

Algorithms and feature expansion

Algorithms:

- Gradient Boosting Machines
- Random Forests
- Linear Classifiers
  - Logistic Regression
  - Support Vector Machines
- Multilayer Perceptron (MLP)

Training sets:

- $\mathcal{D}_0$ the original training set.
- $\mathcal{D}_1$ contains $\mathcal{D}_0$ and the product of every pair of features.
- $\mathcal{D}_2$ contains $\mathcal{D}_0$ and the ratios of every pair of features.
- $\mathcal{D}_3$ contains $\mathcal{D}_0$ and the subtraction of every pair of features.
- $\mathcal{D}_4$ contains $\mathcal{D}_0$ and the new features of $\mathcal{D}_1$ and $\mathcal{D}_2$.
- $\mathcal{D}_5$ contains $\mathcal{D}_0$ and the new features of $\mathcal{D}_1$ and $\mathcal{D}_3$.
- $\mathcal{D}_6$ contains $\mathcal{D}_0$ and the new features of $\mathcal{D}_1$, $\mathcal{D}_2$ and $\mathcal{D}_3$.

Results

The test set contained 550000 events, whose labels and weights were unknown.

The leaderboard during the challenge was calculated on 18 % of the test set (public leaderboard).

The final leaderboard was calculated on the other 82 % (private leaderboard).

The strategy was submitting models using different elements in the ensemble and selecting the two best models based on the public leaderboard.

The best solution was obtained using the models obtained using XGBoost and the data sets $\mathcal{D}_1$, $\mathcal{D}_4$,$\mathcal{D}_5$.

This model reached an AMS of 3.75864 in the private leaderboard, finishing 9th of 1785 teams.

Conclusions

- This is a simple but efficient way to deal with the maximization of the AMS.

- Easily parallelizable:
  - The algorithms can run in a multithread environment.
  - Bagging can be implemented in parallel.

- The performance has been tested finishing 9th of 1785 teams.

Acknowledgements

My deepest gratitude:

- To the organizers.
- To everyone who participated in the challenge.

Thank you!!