The Inverse Bagging Algorithm: enriching signal by inverse bootstrap aggregating

Pietro Vischia[1]

[1] Universidá d'Uviéu

UNIVERSIDAD DE OVIEDO

(work in collaboration with T. Dorigo)

XII Quark Confinement and the Hadron Spectrum - QCHS12

The problem

- The most popular classification algorithms require a well modeled signal and a well modeled background
  - For the classifier to learn how to separate the two classes, both models are required
- What if either signal or background has an unknown p.d.f.?
  - Very well known background modeled from simulation, but an unknown signal
  - Very well known signal modeled from simulation, contaminated by a background of origin unclear and/or not simulable
- How to manipulate (enhance/suppress) the fraction of the unknown process, without modifying the kinematic distributions of the very well known one?

Bootstrap: when you don't know the p.d.f....

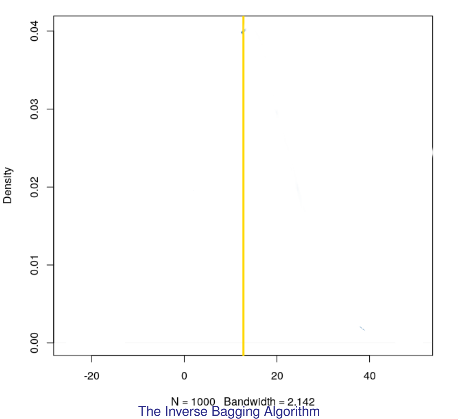
- Observed data $\mathbf{X} = (X_i, ...X_n)$ are sampled from a probability density function F
- Study a statistic of the data, $R(\mathbf{X}, F)$ (e.g. the mean)
- It is often useful to extract its sampling distribution
  - Draw many samples $\mathbf{X}_i$ from the population
  - Compute $R(\mathbf{X}_i, F)$ for each i
  - Study the sampling distribution of the test statistic (e.g. the distribution of the mean)
- Sometimes, we cannot draw additional samples $\mathbf{X}_i$
  - The p.d.f. F underlying the data is unknown
  - We might not have access to the population (e.g. cannot draw more than one sample for 2006 stock market data)
  - It might be unfeasible or expensive (e.g. limited access to telescope time)
- One is left with the single set of sampled data $\mathbf{X}$

... pull more samples from the same data!

- **Plug-in principle:** If the population is not accessible, then sample from an estimate of it
  - Consider your sampled data $\mathbf{X}$ as an estimate of the population, $\hat{F}$
  - Draw many samples $\mathbf{X}^*{}_i$ from $\mathbf{X}$ with replacement
  - Compute $R^*(\mathbf{X}^*{}_i, \hat{F})$ for each i
  - Study the bootstrap distribution of the test statistic
- Key concept: "sampling with replacement".
  - Sample two with replacement from $\mathbf{X} = \{A, B, C, D, E\}$
  - Pick C ($p = 1/5$), put back C, pick A ($p = 1/5$), put back A, pick B ($p = 1/5$). Sample is $\{A, B, C\}$
  - Pick E ($p = 1/5$), put back E, pick C ($p = 1/5$), put back C, pick E ($p = 1/5$). Sample is $\{E, E, C\}$
  - The samples are independent (covariance zero)
- When you sample without replacement, covariance is $-\frac{\sigma^2_{pop}}{N_{pop}-1}$
  - Pick C ($p = 1/5$), pick A ($p = 1/4$), pick B ($p = 1/3$). At any pick you cannot pick the previous picked one

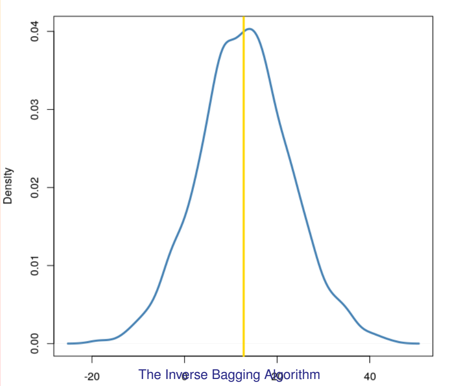
A practical example - 1

- Take financial data
  - Vector **X** of daily returns x_i of IBM for the year 2006, from http://www.burns-stat.com/pages/Tutor/spx_ibm.txt
- Statistic: yearly return $R(\mathbf{X}, F) = \sum x_i$
- F is unknown
- Cannot resample (we cannot "replay the year 2006")
- We have only one value for the statistic
- How can we estimate its variance?



A practical example - 2

- Take financial data
  - Vector **X** of daily returns x_i of IBM for the year 2006, from http://www.burns-stat.com/pages/Tutor/spx_ibm.txt
- Statistic: yearly return $R(\mathbf{X}, F) = \sum x_i$
- F is unknown
- Cannot resample (we cannot "replay the year 2006")
- We have only one value for the statistic
- Draw 1000 samples with replacement from the set of yearly samples
- You can now estimate the variance from the bootstrap distribution



A step forward: B(ootstrap)Agg(regat)ing

UNIVERSIDAD DE OVIEDO

- A multivariate classifier is a statistic of the data
  - The specific form of the function is chosen by some optimization criteria on a training sample
  - Being a statistic, it is open to be estimated via bootstrapping!
- **Boostrap aggregating:** apply a given classification technique to many training sets obtained via bootstrap
  - Obtain many independent classifiers
  - For each event, its final classification is a majority vote between the individual classifications
  - General procedure, applicable to nearly every classification technique
- Main benefit: classification is less dependent on statistical fluctuations in the training sample
  - It can significantly improve classification performance: it can outperform even boosting techniques - Ilya Narsky ("Optimization of signal significance by bagging decision trees", arXiv:physics/0507157, 2005)

Inverse bagging - building subsamples

UNIVERSIDAD DE OVIEDO

- Suppose to be in the "very well known background, unknown signal" case
- Start from a test sample of N_{test} events
  - Suppose it is constituted by 90% background events and 10% signal events
- Take a training sample of N_{train} background events
  - We trust our simulation to be modeling background very well
- Let's randomly pick up from the test sample $M << N$ events
  - On average, $0.9 \times M$ events will be background events
  - With some fluctuations
  - There is a small chance that all M events are from background
- Compare the features of the M events with the background p.d.f. from the training sample
  - Statistical test answering to the question "how likely is that the M-events set is background-like?"

Inverse bagging - basic majority vote

- Using bootstrap, choose a very large number of subsets of M events
  - One can have as many subsets classified as background-rich as desired
- **For each test event i, count how many times it is picked up in a background-like subset ("$ok[i]$")**
  - Weight that number by the number of times the event was picked up to be part of a subset ("$tried[i]$")
- To evaluate performance in terms of efficiency and purity, remove progressively events with largest ok/tried ratio
- ok/tried ratio can be substituted by the average value of the test statistic over the subsets

Compare *M*-subset of test sample with MC training sample

UNIVERSIDAD DE OVIEDO

- *M* is small $\rightarrow$ cannot rely on χ^2
- Kolmogorov-Smirnov test statistic
  - But many variables in HEP use cases differ mainly in the tails
- Anderson-Darling test statistic
  - Designed to be more sensitive to the tails of the distributions
- Energy test (Zech)
  - Multi-dimensional, based on weighted distances
- Personalized multi-dimensional GoF test
  - Based on nearest-neighbour distances ratio R, but use Zech's approach (potential energy of set of charges of magnitude R)
  - Enhanced power for testing localized differences between the distributions
  - Computationally costly (slow)

How to test performance with respect to the market

UNIVERSIDAD DE OVIEDO

- A meaningful comparison requires using MVA methods that do not rely on the p.d.f. of the non-well-modeled sample
- Relative Likelihood (discriminating power from ratio b/ween p.d.f. of the test and of the training sample)
- kNearest-Neighbour (discriminating power from ratio b/ween integrated distance of test event from background events and of test event from other test events)
- Both these reference methods use **event based variables**, whereas inverse bagging uses subset properties to infer event classification
  - Open question: is there any a-priori proof that a sample-based statistic cannot contain more information than an event-based statistic?
  - Oper question, rephrased: is there any a-priori limit to the amount of information that can be extracted using a sample-based statistic? Is this limit related to the amount of information that can be extracted using an event-based statistic?

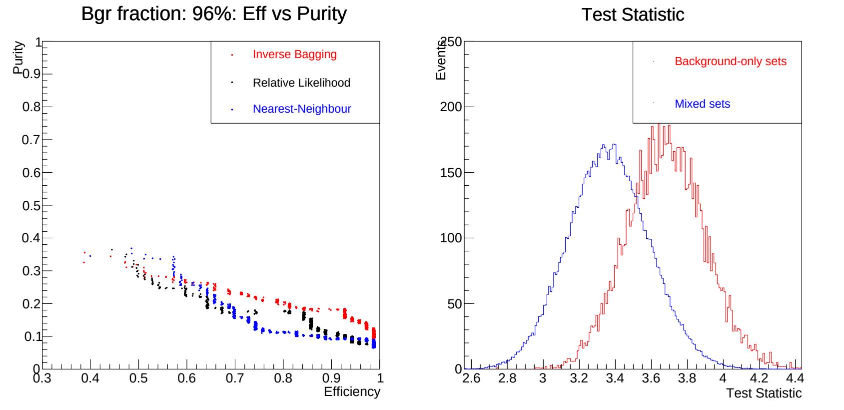
Test performance

UNIVERSIDAD DE OVIEDO

- Use the HEPMASS (http://archive.ics.uci.edu/ml/datasets/HEPMASS) dataset
  - P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, and D. Whiteson. “Parameterized Machine Learning for High-Energy Physics.”, http://arxiv.org/abs/1601.07913v1
- Background: simulated $t\bar{t}$ events decaying semileptonically
- Signal: simulated new particle X with $M_X = 1000\ GeV$, decaying into $t\bar{t}$ pairs
- We know very well the $t\bar{t}$ kinematics: in case of the signal, the intermediate resonance will modify its kinematics
  - Let’s assume we don’t know this signal, and use it to populate our test sample
- The full dataset provides low-level variables (lepton and jets four-momenta, b-tagging discriminators...) and high-level variables ($M_{\ell\nu}$, M_{WWbb}...)
- Test run with 8 low-level variables not including b-tagging discriminators

A basic proof of concept

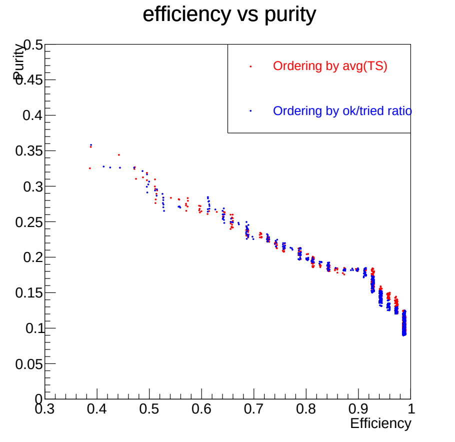
- Pure background training set: 5000 events
- Test set: 1000 events (background fraction: 93%)
- Bootstrap: 100k subsets with 100 events each
- Test statistic: multi-D GoF test
- Rank events by the average value of the TS on each bootstrap sample
- Tested against Relative Likelihood, and k-Nearest-Neighbour
- Inverse bagging outperforms both, particularly for high efficiencies

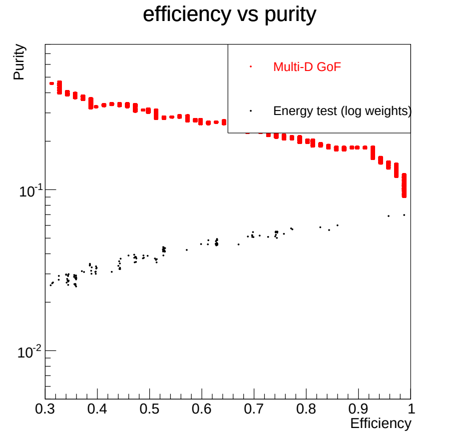


A basic proof of concept - ordering principle

UNIVERSIDAD DE OVIEDO

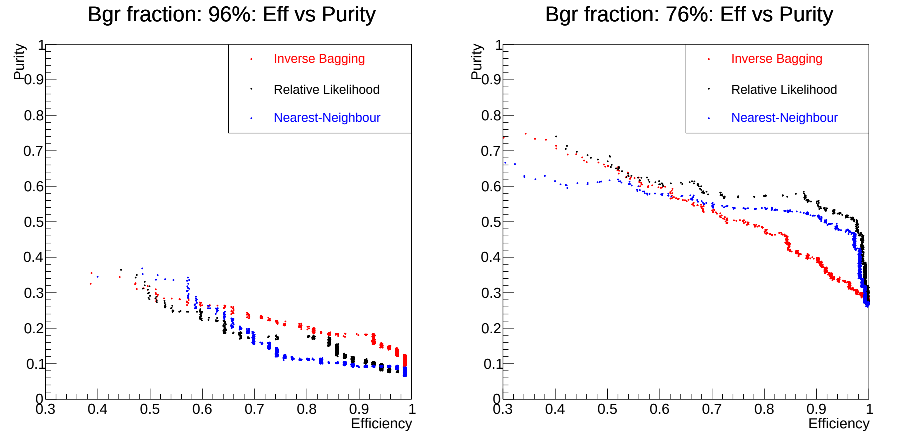
- Ranking by the probability of inclusion yields a similar performance than the more sophisiticated average value of the test statistic
- The algorithm is sensitive to a change in test statistic
    - This needs to be investigated on different datasets and configurations





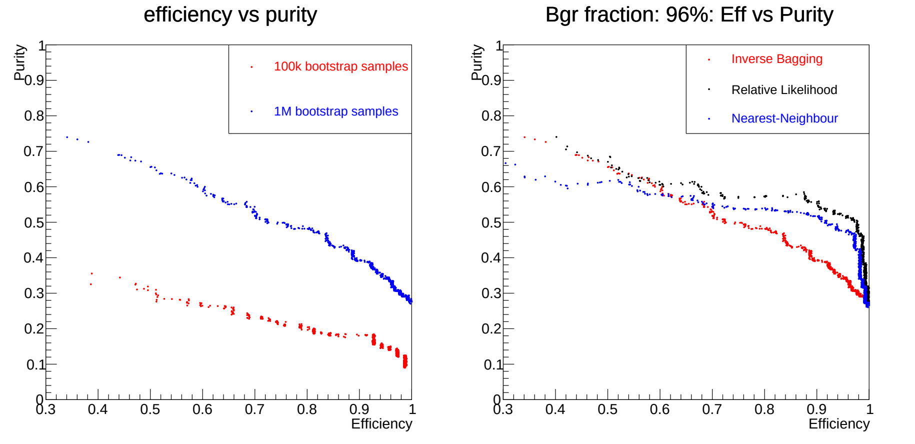
A basic proof of concept - background fraction

- As expected, when the signal fraction is no longer very small, it is more difficult to pick background-like subsets
- The performance relative to NN and RL decreases, since the assumptions behind the algorithm no longer hold
- The other classifiers, as expected, become far better in classification, having more signal events to pick features from



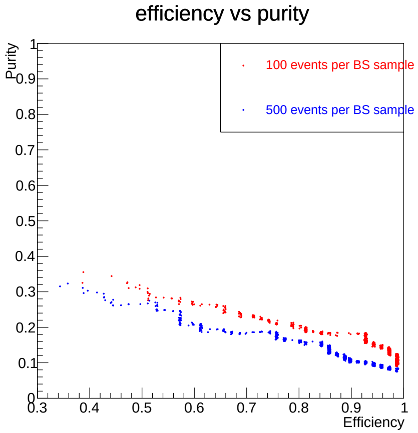
A basic proof of concept - number of bootstrap samples

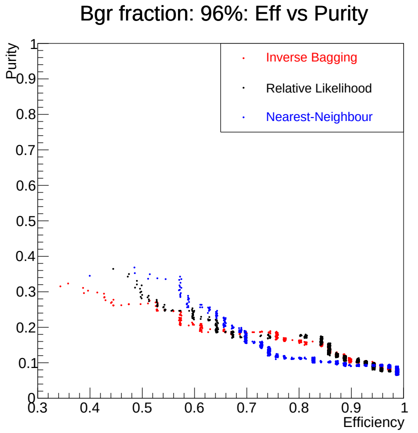
- Increasing the number of bootstrap samples has a high impact as well
- The number of bootstrap samples cannot be too large, otherwise performance is lost w.r.t. benchmark classifiers



A basic proof of concept - size of each bootstrap sample

- Increasing the size of each bootstrap sample can worsen the performance





Summary and perspectives

UNIVERSIDAD DE OVIEDO

- Bootstrap is a powerful method useful in cases where the underlying p.d.f. is not accessible
- Aggregating the information coming from bootstrap enables building multivariate classifiers known to outperform the basic ones
- When a large very well known background is superimposed to an unknown small signal, the *inverse bagging* algorithm can outperform comparable algorithms
  - Situations like this are more and more typical in LHC new physics searches
  - Useful for anomaly/outlier detection!
- Encouraging results with HEPMASS dataset how the feasibility of this algorithm
  - Open theoretical question (sample-based statistic vs event-based statistic)
- The choice of test statistic is an important parameter of the algorithm
  - More tests are ongoing by varying the bootstrap sample size as a function of the number of bootstrap samples generated
  - MultiD-GoF outperforms other test statistics (KS, AD, Zech's ET)

THANKS FOR THE ATTENTION!

Backup