TOP TAGGING USING SPATIAL DISTRIBUTION OF SUBJECTS

Danielle Wilson
Supervisor: Professor Deepak Kar

University of the Witwatersrand

High Energy Particle Physics (HEPP) Workshop 2020

January 29, 2020
Introduction
  ▶ Jets
  ▶ Boosted Objects
  ▶ Top Jets
  ▶ QCD Jets

Classifying Events Spatially
  ▶ Notation
  ▶ Identifying Configurations of Subjets
  ▶ Probabilities of Configurations

Creating the Tagger
  ▶ Combining Probabilities
  ▶ How did it perform?
  ▶ Shortcomings of the Tagger
INTRODUCTION
Quarks and gluons do not occur freely in nature
  ➤ Immediately after production, they fragment and hadronise
    ⇒ collimated shower of energetic hadrons which is referred to as a jet

Can identify original "parton" by measuring jet energy and direction
  ➤ Concept of "parton" is ambiguous
    ⇒ Jets must be well defined
      • Jets defined by algorithm used to assemble them and a radius parameter
      • No single universal definition

Majority of events at the LHC contain jets
Boosted Objects

Figure: Illustration of the decay of a boosted top quark. Ref: arXiv:1712.01391

- If heavy particles are produced with large transverse momentum, decay products collimated into a single large-radius jet.
- Large-radius jets contain intricate substructure.
  - Observables are constructed to characterise this substructure.
- Using one or more of these observables to identify boosted objects is referred to as tagging.
Top quark has a very short lifetime
\[ \Rightarrow \] decays before it hadronises
\[ \Rightarrow \] unique opportunity to study bare quarks

Top quarks decay mainly via
\[ t \rightarrow Wb \]
- \[ W \rightarrow q\bar{q} \] occurs 67% of the time
- \[ W \rightarrow l\nu \] has a branching ratio of 11% for each lepton flavor
- Pairs of top quarks: 45% hadronic, 35% semileptonic, rest are dileptonic and hadronic tau decays
Standard methods to identify Top Quarks:
- B-tagging
- Identifying W boson
- Invariant mass of 3 jets is comparable to the top mass

Highly boosted top quarks
⇒ Standard methods are hindered
⇒ Jet substructure analysis is the natural next step
Efficient top taggers discriminate features unique to top quarks from those of the background.

QCD jets describe the background:

- They originate from high $p_T$ light quarks or gluons that shower into many soft and collinear particles.

$\Rightarrow$ Not easily resolved.
Classifying Events Spatially
**NOTATION**

**Figure:** Illustration of the notation used in this investigation

- Plots in the eta-phi plane
- Large radius jets have a radius of 1, whilst the subjets had radius 0.2.
- Only events with more than 2 subjets were considered.
In order to consistently classify configurations of subjects, a clustering algorithm was implemented. K-means clustering algorithm was chosen for this analysis:

- Separates data into K pre-defined clusters
- Clusters do not overlap
- Aims to maximize similarity between cluster points and the distance between clusters.
- It is easy to implement
Problem: pre-defining number of clusters

⇒ Silhouette Analysis applied

▶ It determines optimal number of clusters ⇒ greatest separation between clusters

▶ Silhouette score ∈ [−1, 1] assigned to measure degree of separability

▶ 1 ⇒ very good clustering
-1 ⇒ very bad clustering

Problem: Not possible to define 1 cluster or for clusters to have single data points

⇒ distance cuts applied
**Figure:** Plot comparing the probabilities of different spatial configurations for an event containing 3 subjects
**Figure:** Plot comparing the probabilities of different spatial configurations for an event containing 4 subjects.
Creating the Tagger
Two variables from events used to create final tagger:
- Number of Subjets
- Spatial configuration of Subjets

Is the event more likely to be signal or background?
⇒ Implementation of Naive Bayes Classifier
- It combines probabilities of the two variables from event data and previously determined probabilities from “training” data
Not well..

- **Signal Efficiency**: $\epsilon_s = 47.1\%$
- **Background Rejection**: $1 - \epsilon_B = 50.6\%$

**BUT** creating a super efficient tagger was not the purpose of the project

**Analysis was too simple to obtain viable results**

**Important qualitative results**

- **QCD Jets**: subjets tended to be closer together
  - 3 Subjets: ‘123’
- **Top Jets**: subjets tended to be more distinct
  - 3 Subjets: ‘1 2 3’
K-means algorithm works best clustering large amounts of data
  ▶ This investigation dealt mainly with only 3 - 6 subjects
K-means initially assigns clusters at random
  ⇒ clustering is not unique
  ▶ Number of events was small
    ⇒ significantly different results obtained each run of the program
The sample of events analysed had unrealistic proportions of Signal to Background events.
The “training” and “testing” data had different proportions of Signal to Background events
  ⇒ Naive Bayes classifier was compromised
Thank you for Listening!