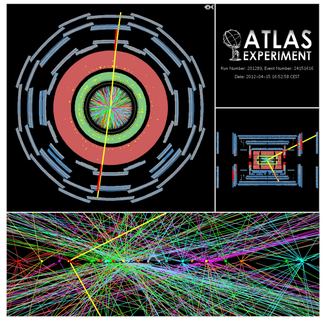
Identification of the correct hard-scatter vertex at the Large Hadron Collider(LHC)

Pratik Kumar(pratikk), Neel Mani Singh(neelmani)

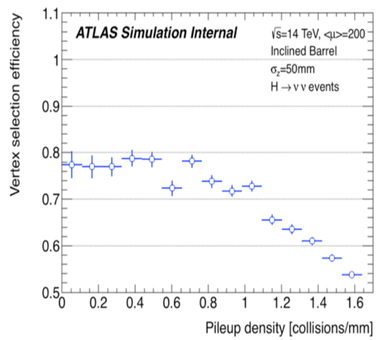
MOTIVATION

- ATLAS is a particle detector analyzing proton-proton collisions from the LHC.
- Identification of the correct hard-scatter primary vertex from around 60 collisions.
- Key challenge for the analysis of LHC events is pileup.



CURRENT METHOD

The current technique for the identification of the primary vertex selects the vertex with the highest total energy. The total energy is computed as the scalar sum of all particle tracks associated to the vertex. This method has a very poor performance when the number of pileup interactions is large, selecting the wrong vertex 40% of the time as seen in the graph.



DATASET & FEATURES

Our dataset consists of computer simulated events of Higgs bosons. Each event picture consists of a list of vertices (60 on average) and each vertex consists of a list of particle tracks. Each track is represented by a direction in 3D space, an origin (given by the vertex it belongs to), and its energy. res that will be used as inputs for a classifier.

Features used –

- **sumPt** - scalar sum of transverse momentum of all the tracks.
- **sumPtw** - weighted sum of track.
- **MET** - missing transverse energy.
- **eta1, eta2, eta3** - angle for top 3 tracks.
- **pt1, pt2, pt3** – transverse momentum of top 3 tracks.

MODELUSED

- Logistic Regression(LR)
- Neural Network(NN)
- Balanced Bagging(BB)
- Balanced Bagging with Logistic Regression(BBLR)
- Balanced Bagging with Neural Network(BBNN)

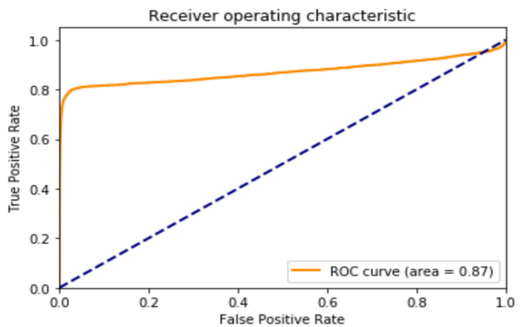
LR did not perform well due to class imbalance in data. Bagging techniques gave better results.

RESULT

Since we have a class imbalance problem, we have to use a metric that is not biased towards the majority class. Therefore we have chosen to use weighted F1-score.

Model	F-Score (test)	F-Score (train)
LR	98.63	98.62
NN	96.84	96.72
BBLR	96.37	96.32
BBNN	55.18	55.01

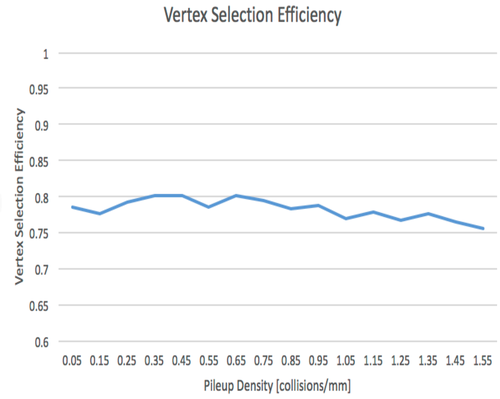
The ROC curve for BBLR indicates that results can be improved by using different threshold



VERTEX SELECTION

- Till now, we have treated each of the vertex as an independent data input.
- But for our problem, we need to select a vertex from a group of vertices of an experiment. For this we evaluate our model per experiment and chose the vertex that gives the highest probabilities.
- Based on this, we calculate the vertex selection efficiency vs pileup densities.

VERTEX SELECTION EFFICIENCY



The vertex selection efficiency from BBLR shows that our model performs better at high pileup densities than the current technique.

DISCUSSIONS

- The data is inherently unbalanced because of the nature of the experiment so general training techniques doesn't work.
- Features apart from sumPt has discriminating effect for different type of collision event. That is why our model works better than the existing approach at high pile-up densities as per vertex selection efficiency.
- Our model performs almost similar on training and test set. Therefore no overfitting.
- Neural Network without balanced bagging method of classification is unstable for this dataset as it produces quite varying results.

FUTURE WORK

- Neural Networks can be improved by tuning of the parameters - learning rate, hidden layer units, etc.
- Thresholds - Predictions based on different thresholds.
- Features - More features can be extracted from the simulation of the events.

REFERENCES

- https://atlas.cern
- Slides from Prof Ariel Schwartzman
- Debashree Devi, Saroj kr. Biswas and Biswajit Purkayastha, "Redundancydriven modified Tomek-link based undersampling: A solution to class imbalance", 2016
- Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall and W. Philip Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", CoRR , 2011.
- https://svds.com/learning-imbalanced-classes/ .