機械学習を用いた CALETイベントデータの識別手法

小門 都澄[1]　古谷 泰愛[1]　吉田 健二[1]

[1]芝浦工業大学大学院 システム理工学専攻

平成29年度 宇宙科学情報解析シンポジウム 2018/2/16

研究背景と目的

背景

- 電子の観測により宇宙線加速源の特定や暗黒物質の解明
- 宇宙線の陽子に対する電子の成分比

  10 GeV : 1%　　100 GeV : 0.3%　　1 TeV : 0.1%
- 電子の観測の際には電子と陽子の識別が必要

- CALETにより今までにない高エネルギー電子領域での宇宙線の観測が可能
- CALETは観測データから粒子の識別を行う必要

目的

CALETにおける電子,陽子の最も識別性能の高い機械学習の提案

CALET CALorimetric Electron Telescope

- 国際宇宙ステーション（Internal Space Station: ISS）の
  日本実験棟「きぼう」に搭載されている宇宙線観測装置

- 2015年10月から運用を開始し，2～5年間の観測を行う予定

図1. 国際宇宙ステーション

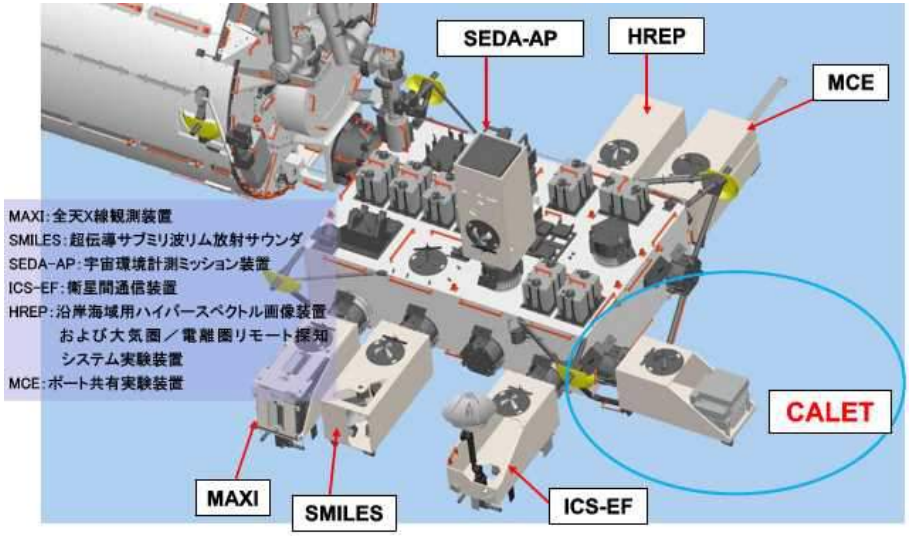


図2. CALET搭載位置

提供：宇宙航空研究開発機構（JAXA），
http://spacestationexplorers.org/lesson-plans/design-a-space-station/

搭載機器 (カロリメータ)

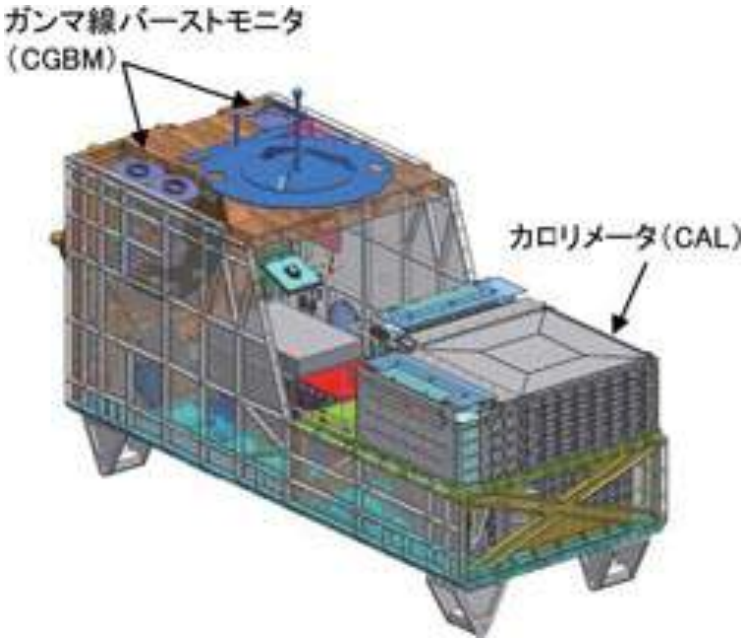


図3.CALET構成機器

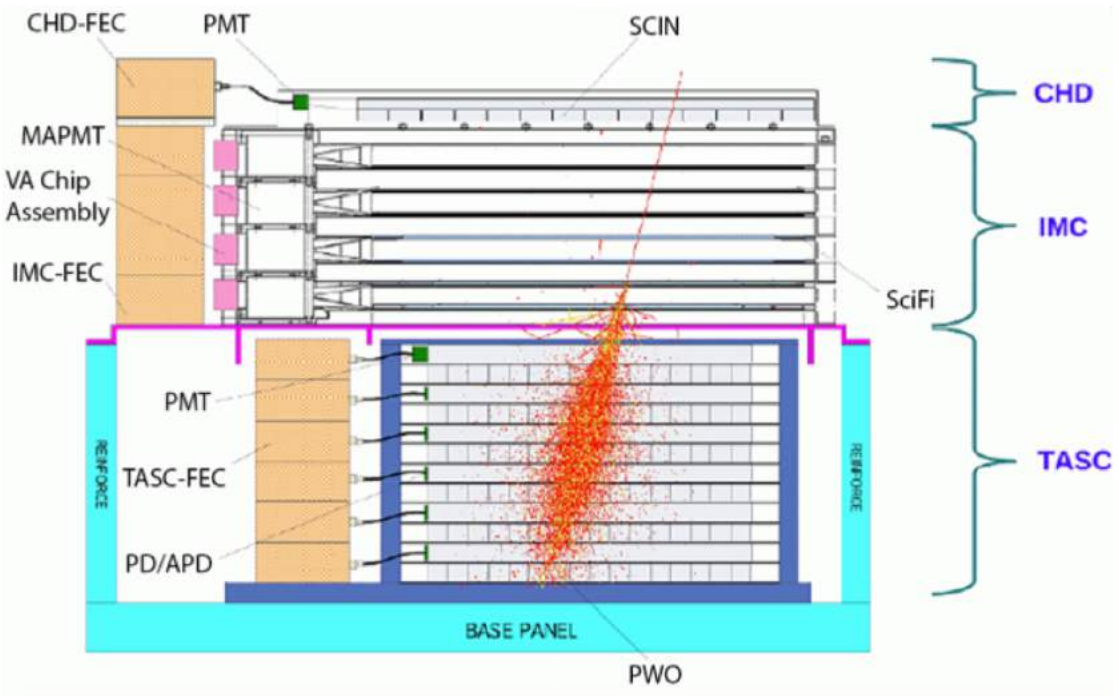


図4.カロリメータの構造とシャワー粒子

- 電荷測定器(CHD):入射粒子の電荷測定を目的
- イメージングカロリメータ(IMC):シャワー初期発達や到来方向の決定を目的
- 全吸収型カロリメータ(TASC):シャワーの発達の様子やエネルギーの測定を目的

提供:宇宙航空研究開発機構(JAXA),
http://calet.jp/id28/

比較する機械学習

6つの機械学習をPythonで実装

- Support Vector Machine (SVM-Linear)
- Support Vector Machine (SVM-RBF)
- ロジスティック回帰
- Boosted Decision Tree (BDT)
- Gradient Boosted Decision Tree (GBDT)
- Deep Neural Network(DNN)

BDTとGBDT

- Boosting

精度の低い分類器(弱分類器)から精度の高い分類器(強分類器)を生成

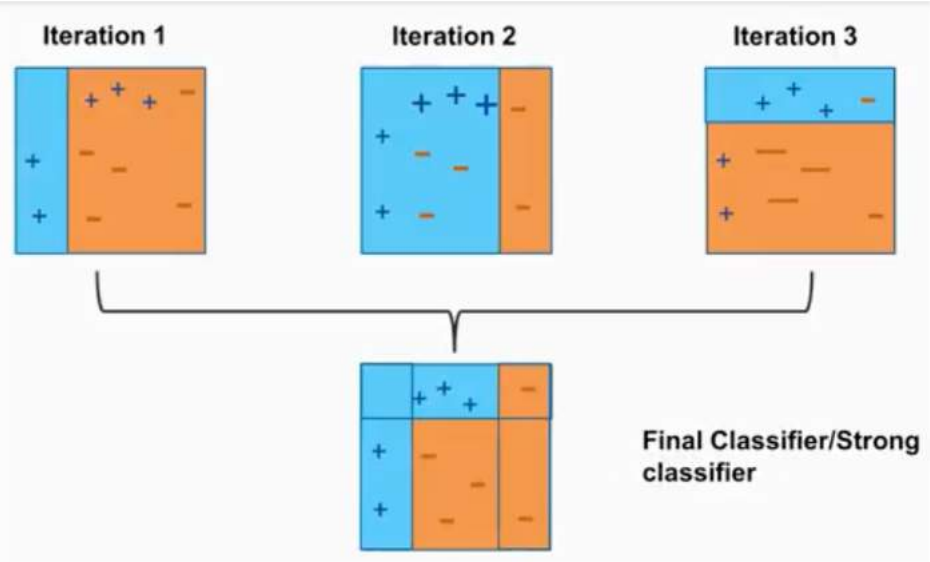


図5. Boostingの手法

- BDT
  - 弱分類器：決定木
  - Boosting：Adaboost

  誤って分類された標本に対応する重みをより重しながら学習器を作る
- GBDT
  - 弱分類器：決定木
  - Boosting：勾配ブースティング

  今ある学習器の結果と目的の値との差(勾配)を縮めるような学習器を作る

参考: https://www.youtube.com/watch?v=BoGNyWW9-mE

Deep Neural Network(DNN)

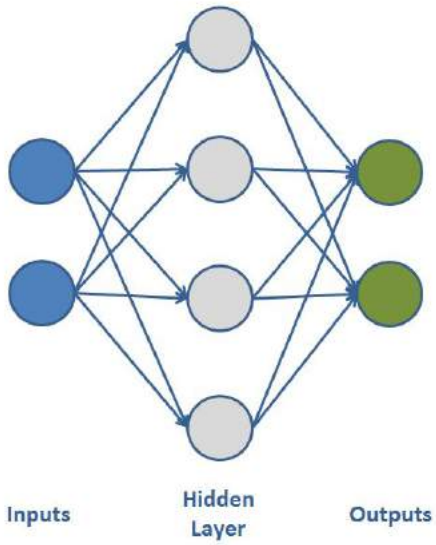


図6.ニューラルネットワーク

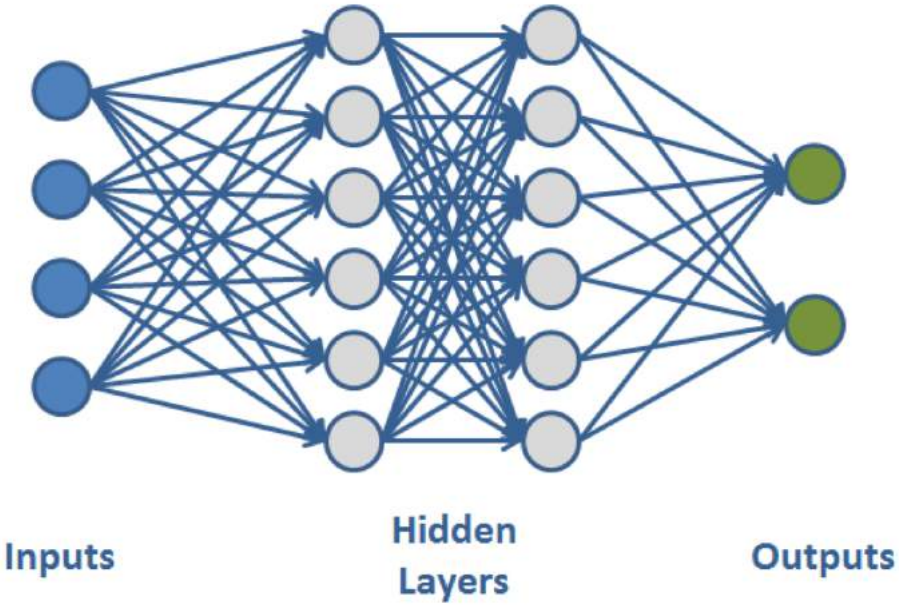


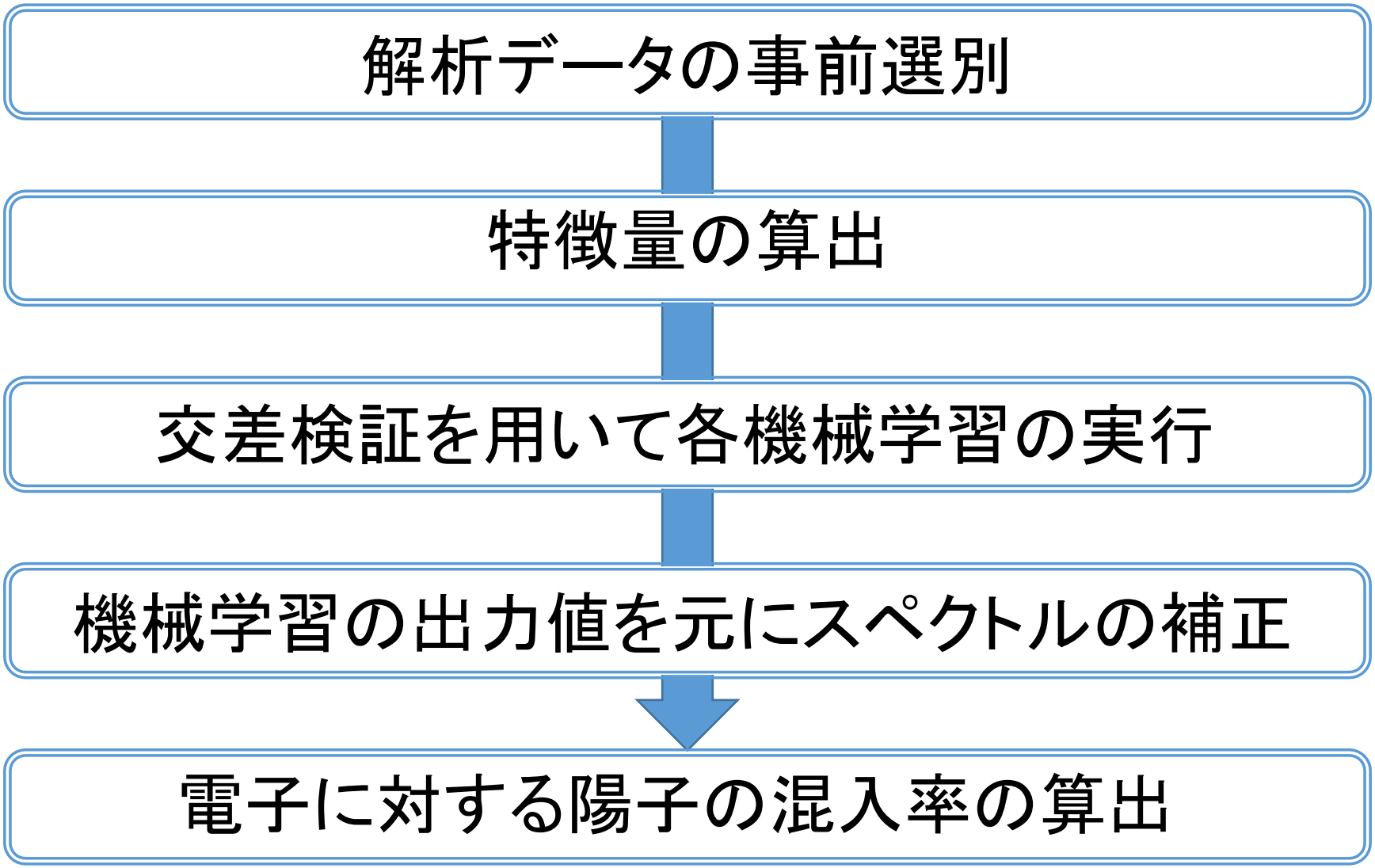
図7. ディープニューラルネットワーク

今回設定したパラメータ

- 入力層：16ユニット
- 中間層：2層, 100ユニット
- 出力層：2ユニット
- 活性化関数：tanh関数
- 最適化手法：SMORMS3

参考: http://davidvandegrift.com/blog?id=47

解析の流れ



事前選別

1. Good EM track

・IMCのXY各軸において8層のうち信号を検出した層が3層以下のイベントを除外

→決定精度などが悪いイベントの除外

2. Reconstructed Geometry

・幾何条件Aのイベントを抽出

→シャワー軸が全ての検出器を十分に通過しているイベントの抽出

3. Offline Shower Trigger

・IMC-X7+X8, Y7+Y8 > 50MIP,

TASC-X1 > 100MIP　※MIPは規格化した粒子数

→高エネルギー電子イベントの抽出

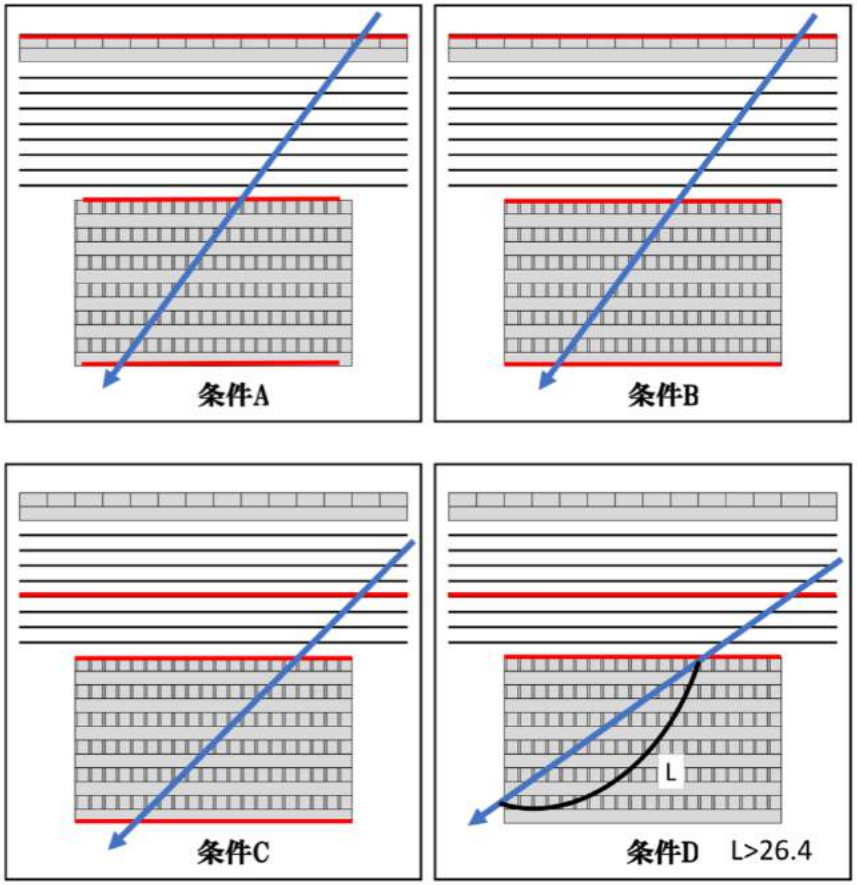


図8. 幾何条件の定義

使用データ

- モンテカルロ(MC)シミュレーションデータ
  - 電子：20 [GeV]～20 [TeV]　スペクトル：E^{-1} (3×10^{7}イベント)
  - 陽子：20 [GeV]～20 [TeV]　スペクトル：E^{-1} (5×10^{8}イベント)
    20[TeV]～1000[TeV] スペクトル：$E^{-2.5}$ (5.8×10^{7}イベント)
- スペクトル補正用データ
  - 電子：AMS-02
  - 陽子：AMS-02 + CREAM

表1.事前選別後の各エネルギー毎のイベント数

エネルギー範囲 [GeV]	91~115	115~144	144~183	183~228	228~290
電子	7570	7180	7642	6896	7586
陽子	14524	14290	15435	14404	15891
エネルギー範囲 [GeV]	290~366	366~459	459~580	580~728	728~920
電子	7105	7086	7269	7070	7218
陽子	15799	15175	15888	14959	15272
エネルギー範囲 [GeV]	920~1152	1152~1459	1459~1845	1845~2308	2308~2912
電子	6970	7278	7179	1520	351
陽子	14708	15189	14804	14113	14486

特徴量(1/2)

電子と陽子では検出器内でのシャワーの発達が異なる→特徴量

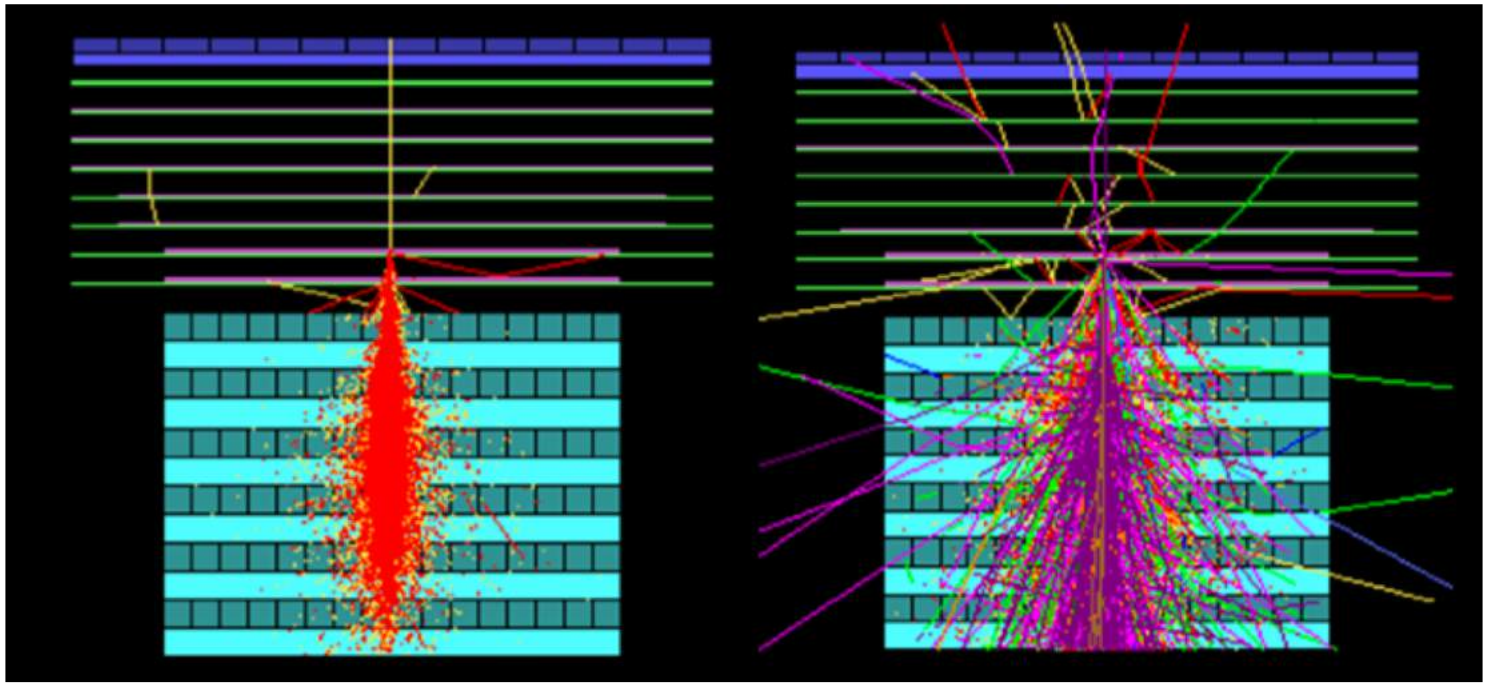


図9. 電子 (左図) と陽子 (右図) のシャワー形状

参考:赤池 陽水 早稲田大学 博士論文 (2012)

特徴量(2/2)

先行研究などを元に16種類に設定

- |(シャワー軸)-(TASC1層目の重心)| : $|\Delta TASC|$
- IMC入射層±1ファイバーのエネルギー損失: IMC_{inc}
- IMC8層のシャワー集中度 : C_{IMC-8}
- CHDのエネルギー損失和 : CHD_{sum}
- TASC Y5,X6,Y6層の各シャワーの縦方向の発達:
  $logEDFY5,\ logEDFX6,\ \ logEDFY6$

  EDF:TASCの全エネルギー損失量に対するTASC各軸各層目のエネルギー損失の割合
- TASC X1,Y1層の各横拡がり: $R_{X1},\ R_{Y1}$
- TASC全体の横拡がり: R_E
- IMCの縦発達のフィッティング係数: $p0,\ p1,\ \chi^2$
- TASCの縦発達のフィッティング係数: $T_{max},\ b,\ \chi^2$

16種類

評価方法

- 交差検証 (クロスバリデーション)

1. データを5個に分割
2. 4 個で学習を行い残りの1個でテスト
3. 5 回繰り返す

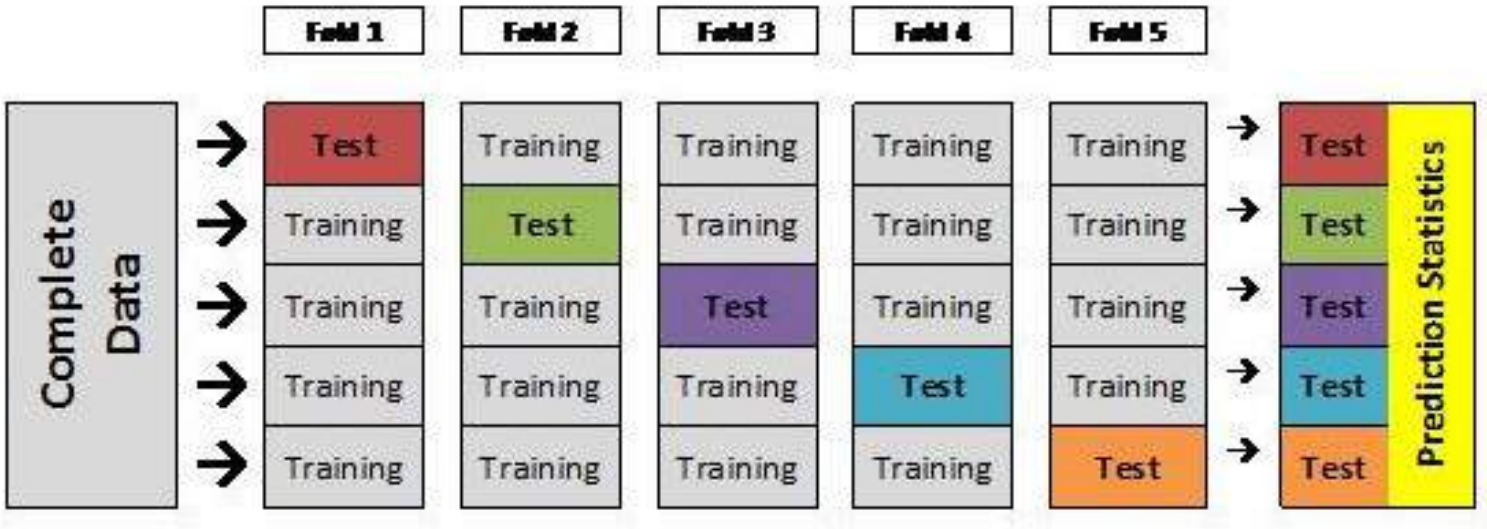


図10. クロスバリデーションの手順 (*K*=5)

交差検証を行うことですべてのシミュレーションデータを検証に使用

参考: http://blog.goldenhelix.com/goldenadmin/cross-validation-for-genomic-prediction-in-svs/

結果(SVM-Liner, SVM-RBF)

- エネルギー範囲：920 [GeV]～1152 [GeV]
- 青：電子　オレンジ：陽子
- 横軸：各機械学習の出力値　縦軸：ビン毎のイベント数

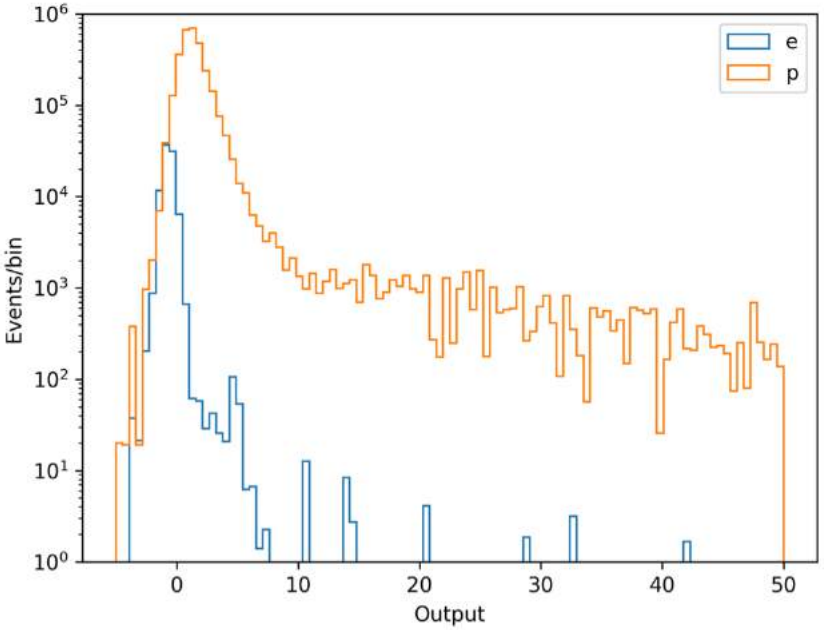


図11. SVM-Linerの出力値のヒストグラム

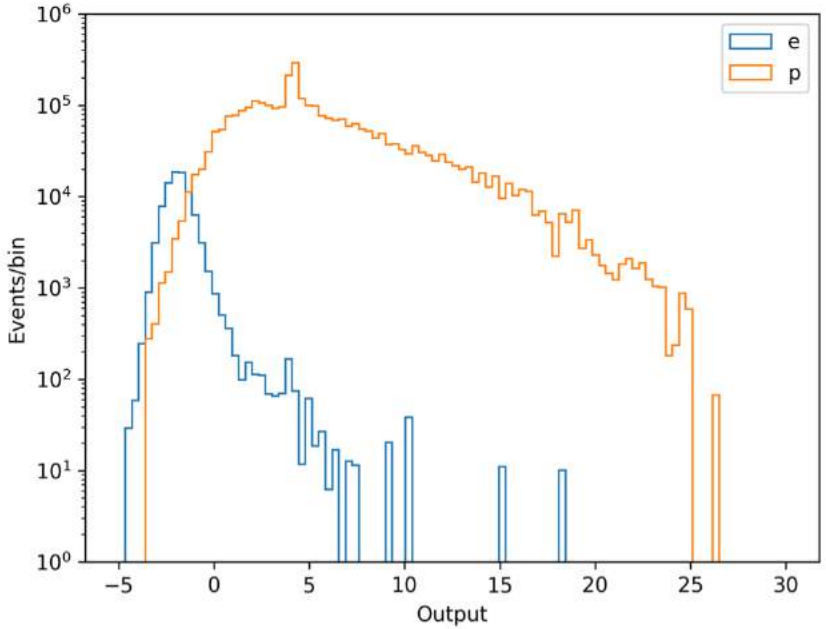


図12. SVM-RBFの出力値のヒストグラム

結果(ロジスティック回帰, BDT)

- エネルギー範囲：920 [GeV]～1152 [GeV]
- 青：電子　オレンジ：陽子
- 横軸：各機械学習の出力値　縦軸：ビン毎のイベント数

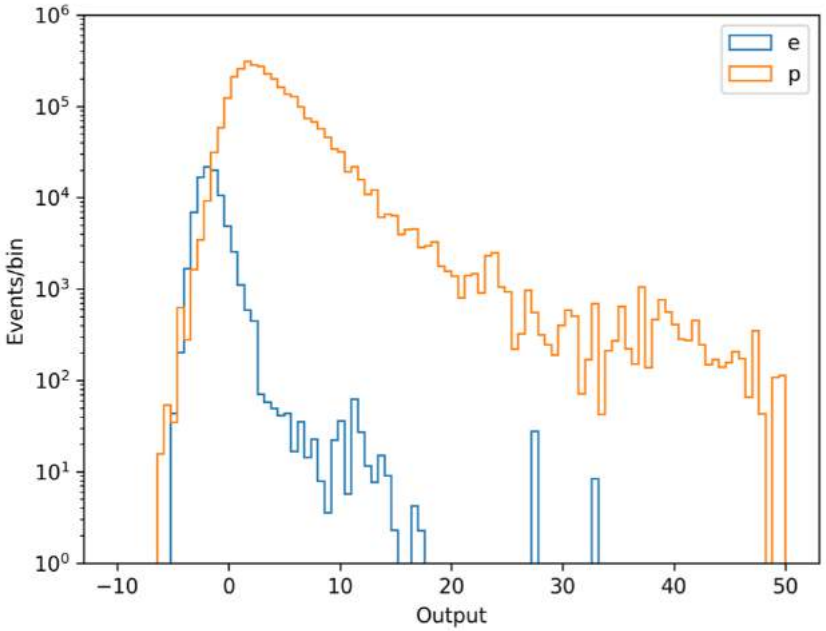


図13. ロジスティック回帰の出力値のヒストグラム

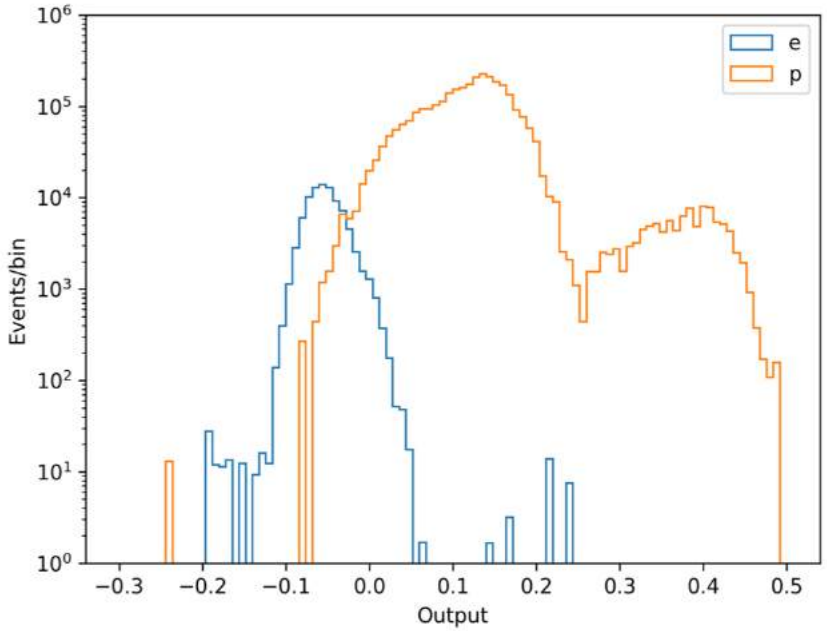


図14. BDTの出力値のヒストグラム

結果(GBDT,DNN)

- エネルギー範囲：920 [GeV]～1152 [GeV]
- 青：電子　オレンジ：陽子
- 横軸：各機械学習の出力値　縦軸：ビン毎のイベント数

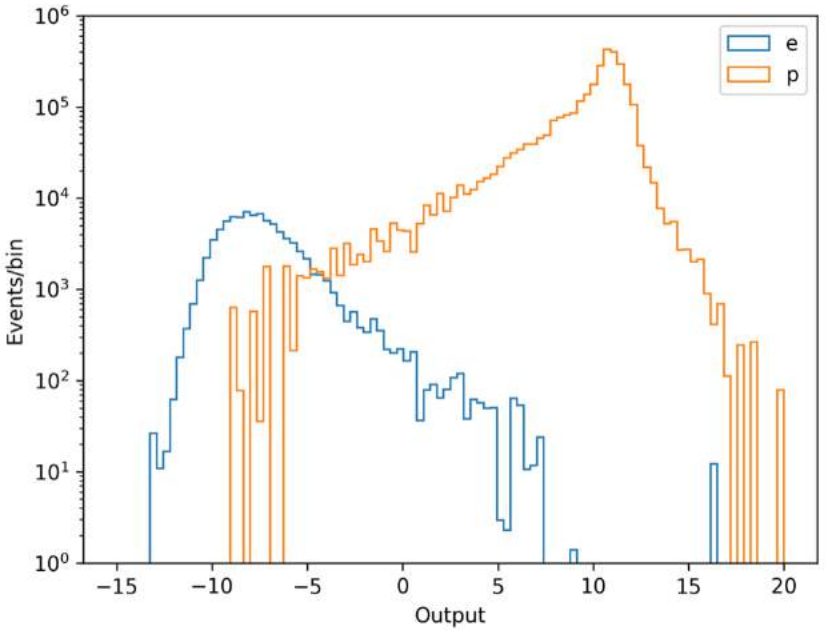


図15. GBDTの出力値のヒストグラム

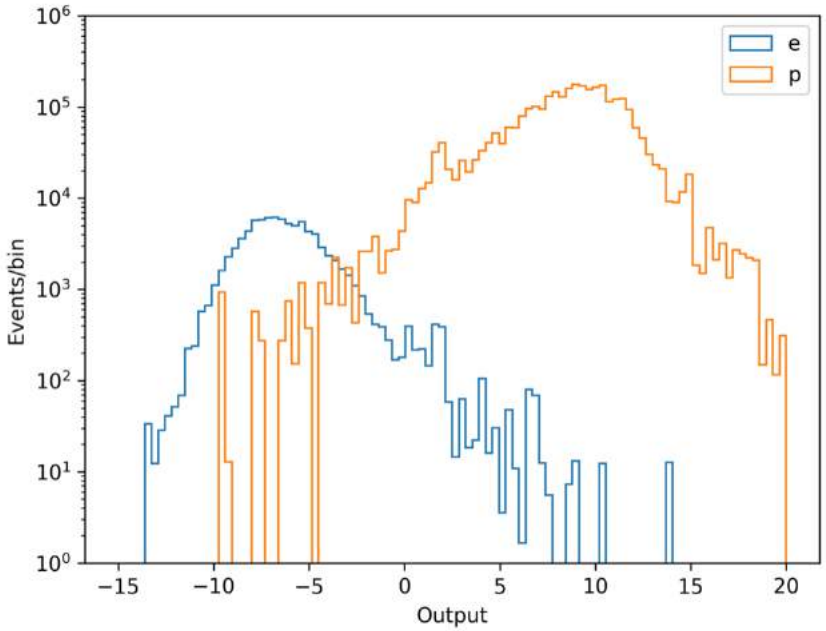


図16. DNNの出力値のヒストグラム

結果のまとめ

- 電子の検出効率を80%にするような閾値をそれぞれ設定
- 評価基準 → 陽子混入率：陽子/(電子+陽子)

表2. 920～1152[GeV]の各機械学習の陽子混入率

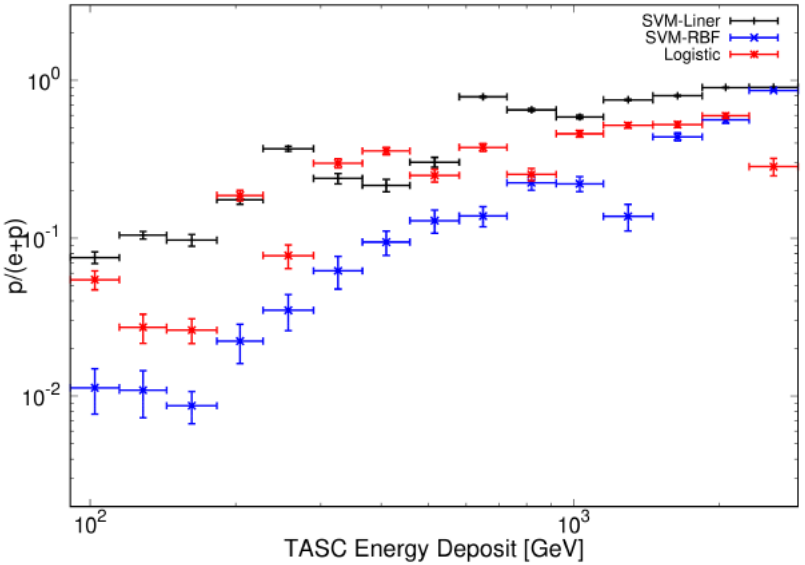
	SVM (Linear)	SVM (RBF)	ロジスティック回帰	BDT (先行研究手法)	GBDT	DNN
閾値	-0.269	-1.246	-0.828	-0.034	-5.760	-4.492
電子検出効率	80.4%	81.1%	80.6%	80.9%	80.9%	81.1%
陽子混入率	0.586 ±0.016	0.221 ±0.024	0.459 ±0.022	0.111 ±0.025	0.064 ±0.019	0.059 ±0.019

考察 (各SVMとロジスティック回帰の比較)

- SVM-Linear, ロジスティック回帰 → 線形分類
- SVM-RBF → 非線形分類

<u>非線形の方が陽子混入率が低い</u>

識別を行う際に引く超平面は，単純な直線では電子と陽子の分離が難しい(非線形をしている)



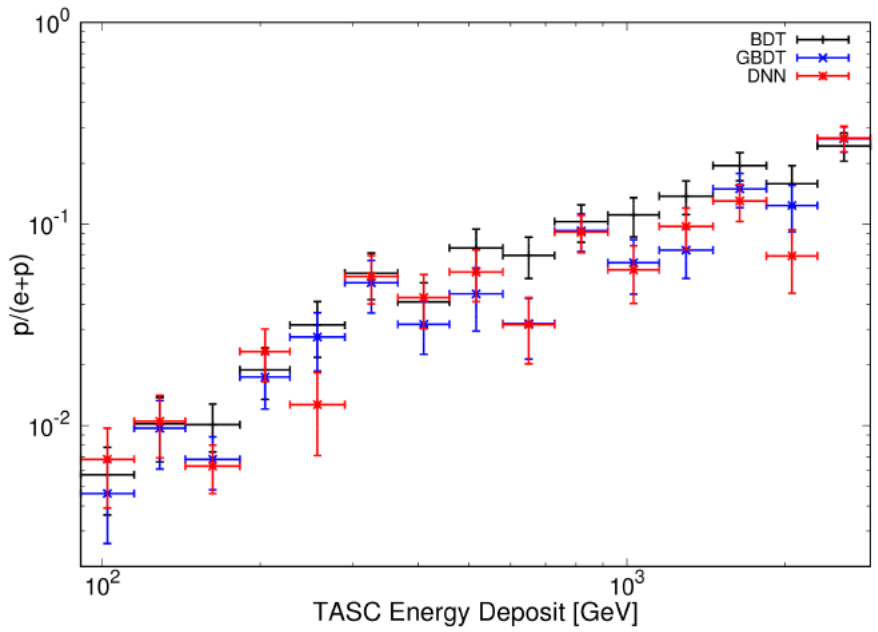
横軸：TASCのエネルギー損失
縦軸：陽子混入率

黒：SVM-Liner
青：SVM-RBF
赤：ロジスティック回帰

図19. SVM-Linear,SVM-RBF及びロジスティック回帰の91~2912 [GeV]における陽子混入率のエネルギー変化

考察 (BDT,GBDT,DNNの比較)

- GBDTの陽子混入率が低い理由
  - 勾配ブースティングがAdaboostよりも性能が高い
- DNNが陽子混入率が低い理由
  - 中間層で有効な特徴量が生成できた



横軸：TASCのエネルギー損失
縦軸：陽子混入率

黒：BDT
青：GBDT
赤：DNN

図20. BDT，GBDT及びDNNの91~2912 [GeV]における陽子混入率のエネルギー変化

結論と課題

- 先行研究の手法であるBDT比べるとGBDTやDNNでは全体的にみて陽子混入率を下げることげできた

電子と陽子の識別に有効な手法として

第1にDNN，第2にGBDT

精度の向上には？

- 高エネルギー領域での学習データ数を増やす
- 特徴量の追加
- 機械学習におけるパラメータの設定の調整
  - グリッドサーチなど